Quality Assurance Evaluation of the *Ceriodaphnia dubia* Reproduction Test

Below are the comments of the Stakeholder Committee members on the draft workplan for the *C. dubia* QA evaluation project.

These comments were submitted to SCCWRP and are shared with the public for transparency.

Please note that SCCWRP is not obligated and will not provide detailed responses to individual comments. Instead, the changes will be discussed with the Stakeholder Committee and the Science panel during scheduled public meetings.

The Conceptual Work Plan is a good starting point. Please add in the discussion that the intent of the Work Plan was that it would get more refined in later phases, as it was explained during the March meeting.

Please include these comments in addition to the comments that I had submitted on the PDF version of the Conceptual Work Plan:

1. The comments I made during the last Stakeholder Meeting:
    a. Include a glossary of terms including, but not limiting to clarify the definition of the terms: dose response testing, low comparability, and control variability.
    b. Add temperature of the lab environment to Table 3 because an EPA guidance document stated that temperature can have an effect on the reproduction endpoint for Ceriodaphnia dubia.
    c. Add a timeline figure to the work plan, to track progress, and adjust dates as needed in future iterations of the work plan.
2. The following new comments:
    a. Formatting comment - In the Table of Contents number each section so we can reference other sections by number.
    b. Explain how the 5-step approach on page 3 fits in with the State Water Board contract for the project and the goals of the Study.
    c. When reviewing historical data look for the following:
        1. Did the test include static renewals or was it a flow-through test, or no renewals?
        2. Sample collection - Were the samples grab or 24-hour composite?
        3. Identify reasons why the test results were tossed out, i.e., why the lab deemed the test results invalid
        4. What was the sample holding time
    d. Make the interview mandatory (They can be conducted over the phone or virtually on the computer, if COVID or funding, or other conditions preclude in-person interviews). Interview lab technicians and lab managers, but separately, to get different points of view.

My preference with respect to sequencing would be as follows:

1. Compile exiting data.

2. Review existing data – valid data & tests that were tossed out as not being valid.

3. See if anything stands out, any commonalities, is anything missing from report (identify data gaps).

4. What statistics, calculations, comparisons can be done with the data.

5. Develop a questionnaire based on what was learned from the data review process.

6. Distribute the questionnaire and ask that it be returned by a given date.

7. Review the completed questionnaires.

8. Conduct follow-up interviews (chose best option: phone, virtually, or in person).

9. Conduct lab visit, to see the lab technicians in action (virtual or in-person) look for anomalies.

10. Write a report along the way.

11. conduct split samples/ if time allows.

12. incorporate results of round robin testing into report.

13. Hear from expert panel.

14. finalize report and submit to State Water Board.

With respect to the CASA workplan, please include a set of instructions for the Expert Panel explaining how the attachment came to be, summarizing the discussion that took place during the meeting, and ask them to decide:

1. Does the expert panel find value in conducting split samples & analysis?
2. If so, can they recommend the timing of when that should take place?
    a. Before the historical data is reviewed,
    b. after the historical data is reviewed, or
    c. both times (before and after).
3. Can they come up with a test design to help answer the question of how can people have greater acceptance in the results of a chronic toxicity test?
    a. Identify a set of metrics to measure lab performance and test accuracy.
    b. Identify the factors that can generate different results among 2 labs that analyze the "same water."
    c. Identify standard practices that can improve the way that the test is carried out, i.e., how can the test conditions be optimized.
    d. Recommend the type of information that should be included in a report that conveys the test result .
4. Can the expert panel recommend what elements of the CASA work plan, if any, should be incorporated into the SCCWRP work plan?

I would have appreciated if the CASA document could have included more detail, but I look forward to hearing the recommendations from the expert panel.

If you have questions or comments regarding the draft work plan, please contact your representative on the Stakeholder Advisory Committee:

https://www.waterboards.ca.gov/water_issues/programs/state_implementation_policy/docs/c_dubia_study_stakeholder_meeting_1_stakeholder_committee_members.pdf"

Conceptual Workplan: Quality Assurance Evaluation of the *Ceriodaphnia dubia* Reproduction Test

Prepared by the Southern California Coastal Water Research Project

For the State Water Resources Control Board

Agreement #19-278-0780

Draft - March 18, 2021

Contents

Introduction

The California State Water Board recently adopted Toxicity Provisions, which include numeric effluent limitations for aquatic toxicity. Implementation of the monthly median effluent toxicity limitation for the *Ceriodaphnia dubia* (*C. dubia*) reproduction test has been delayed until January 1, 2024 for some dischargers as specified in the Toxicity Provisions. During this delay, the State Water Board has committed to a study, in collaboration with stakeholders and laboratories, to investigate factors that can lead to test variability. High test variability could lead to low statistical confidence in assessments of toxicity or non-toxicity.

The *C. dubia* test has been used for many decades and the currently promulgated protocol for the *C. dubia* reproduction test was established nearly 20 years ago (USEPA 2002). The promulgated method allows laboratories some flexibility when implementing certain laboratory techniques. For example, there are multiple options for preparing culture water or food. In some instances, the promulgated method is silent on test techniques, leaving laboratories to use their best professional judgement. To be clear, the *C. dubia* method is well-established and validated. However, these small differences between laboratories may lead to intra- or interlaboratory variability, which could influence test results.

Previous studies have assessed the variability of the *C. dubia* test results within and among laboratories. In the early 2000s, an interlaboratory comparison exercise performed by the EPA found that only 22 out 122 *C. dubia* chronic tests did not meet test acceptability criteria for survival or reproduction (USEPA 2001). The invalid tests were confined to 10 out of 34 participating laboratories. The study reported intra- and inter-laboratory coefficients of variation for the IC25 values of effluent and receiving water split samples at 17% and 28%, respectively for the reproduction endpoint. More recently, a smaller intercomparison exercise was conducted in California to evaluate the reliability of *C. dubia* chronic test for stormwater toxicity evaluation (Schiff and Greenstein 2016). Of the nine labs that tested split samples of dilution water, three were considered "low" comparability. Lack of comparability among a minority of laboratories testing split samples of dilution water was also identified by others (Moore et al. 2000; Diamond et al. 2008). Most recently, routine testing data generated by eight California-accredited laboratories was examined by Fox et al. (2019) and results indicated that intra-laboratory variability, particularly in controls, influenced whether test samples would be identified as toxic. Either reducing the between replicate variability or increasing the number of replicates improved lab performance.

The State of California Environmental Laboratory Accreditation Program (ELAP) accredits all laboratories conducting analysis for regulatory compliance purposes, including the *C. dubia* test. Currently, there are 17 ELAP accredited laboratories for conducting the *C. dubia* test in California (Appendix A). Accreditation is based on the demonstration that laboratories are following the testing protocols, properly training their staff, keeping accurate records, and demonstrating they can meet data quality objectives for reference toxicant and performance evaluation test samples. While this process demonstrates that a laboratory capably performs a test, it does not address test variability between laboratories or differences in lab techniques that are allowed by the protocols.

Objective of this study

The objective of this study is to evaluate sources of variability in the *C. dubia* reproduction test conducted by California-accredited laboratories and identify potential laboratory technique guidance and/or recommendations to: (a) improve the consistency of the execution of the C. dubia test method

to achieve improved precision (i.e., as measured by the control coefficient of variation) within each testing laboratory; and (b) improve the consistency and comparability of *C. dubia* test results among testing laboratories.

The study will seek to answer the following questions:

1) What are the *C. dubia* chronic reproduction toxicity test laboratory techniques used by Environmental Laboratory Accreditation Program (ELAP) accredited laboratories in the state of California?
2) How does variability in control reproduction and/or reference toxicant response in the *C. dubia* chronic reproduction toxicity test compare amongst laboratory technique differences used by ELAP accredited laboratories?
3) Does standardizing differences in the *C. dubia* chronic reproduction toxicity test laboratory techniques reduce variability in control reproduction and/or reference toxicant response?

Based on the results of this study, a list of suggested best-practices for the *C. dubia* reproduction test laboratory techniques will be developed.

Just as importantly, this study is not designed to address or quantify false negative or false positive rates for detecting toxicity from known or unknown samples.

Approach

A five-step design will be used to address the study objectives:

1) Create a Governance structure
2) Analyses of historical data and lab techniques provided by ELAP-accredited laboratories to identify sources of variability
3) Dose-response testing to optimize lab technique(s) and recommended lab technique guidance
4) Evaluation of the revised lab technique guidance via split-sample testing by accredited laboratories
5) Final report with final recommended guidance

These steps are sequential with each one informing the details of the next.

The first step will create a two-tiered governance structure to ensure transparency and technical rigor. One tier will be a Stakeholder Committee comprised of representatives from sectors potentially impacted by the study results. The second tier will be an independent Expert Science Panel comprised of scientists with different areas of expertise and no potential conflict with study results.

The second step will be comprised of two subtasks. The first subtask will create an inventory of lab techniques used by ELAP accredited laboratories. The inventory will elucidate the level of comparability and differences in test implementation. The second subtask will collect historical testing data from the ELAP accredited laboratories. The historical data will be analyzed to quantify the level of variability within and among laboratories. Finally, the differences in lab techniques will be compared to the lab test result variability. The goal of this data analysis is to indicate which lab techniques might be accounting for the observed variability. An optional subtask is to collect new data to assess intra- or interlaboratory variability using split samples, to confirm possible sources of test variation that historical data does not provide. This split sample testing will be dependent on the availability of additional funds.

The third step will focus on dose-response testing procedures to quantify the variability of lab techniques identified by the historical analysis in Step 2. There may potentially be many variable-inducing differences in lab techniques from Step 2. So, a prioritization of which techniques require dose-response testing will need to be evaluated by the Stakeholder Committee and Expert Science Panel formed in Step 1. The results of the dose-response testing will culminate in a draft recommendation on the lab technique guidance that produces the least variability in test results.

The fourth step will verify that draft recommendation from Step 3 does reduce variability, both within and among laboratories. To accomplish this verification, split samples will be distributed to ELAP accredited laboratories for testing, using the draft recommended guidance for laboratories to follow.

The fifth step will complete the study including final recommended guidance, results from the split sample testing in Step 4, and a final report.

Detailed Methods

Create a governance structure

This task will create a multi-tiered governance structure to ensure transparency and technical rigor. Figure 1 illustrates the governance structure.

The ultimate decision-making body is the State Water Resources Control Board (SWRCB). The SWRCB staff is charged with making the final recommendations to the SWRCB about the need for implementing any recommended guidance on lab technique for the *C. dubia* reproduction test.

The project facilitator is the Southern California Coastal Water Research Project (SCCWRP). SCCWRP will be responsible for project design (including this workplan), project implementation (including interacting with ELAP accredited laboratories), and project reporting (including the final report).

One tier will be a Stakeholder Committee comprised of representatives from sectors potentially impacted by the study results. A list of the Stakeholder Committee sectors, and their designated representatives is included in Table 1. The goal of the Stakeholder Committee is to ensure there is a formal mechanism for input and feedback to the project design, planning, conclusions, and recommendations. While not a decision-making body, the Stakeholder Committee is a crucial piece of governance. The Stakeholder Committee will review any study design, results, and recommended guidance first, prior to the Expert Science Panel, to make sure the study is rooted in applicable and achievable guidance.

The second tier will be an independent Expert Science Panel comprised of scientists with different areas of expertise and no potential conflict with study results. A list of the Expert Science Panel disciplines and the designated scientists are listed in Table 2. The Expert Science Panel is a decision-making body and is tasked with reviewing the study design and approving the Workplan, reviewing intermediate work products and refining the study design, and reviewing the recommended changes to lab techniques and providing a consensus opinion on the final method guidance.

Figure 1. Project governance structure

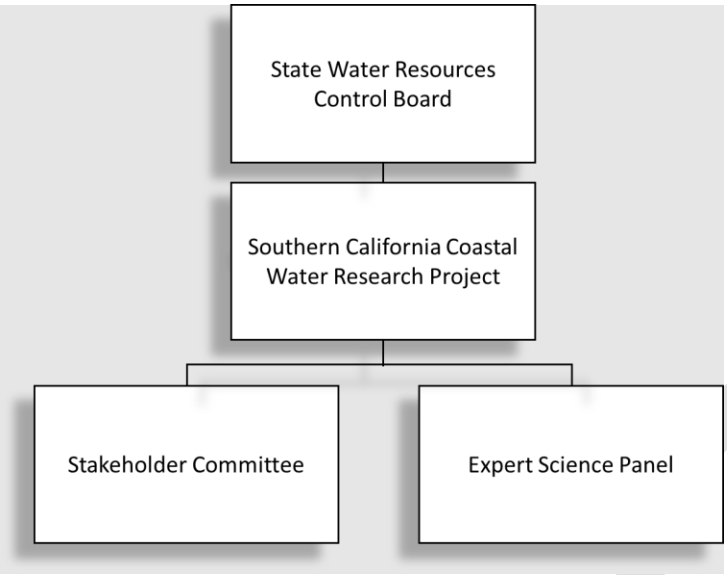


Table 1. List of Stakeholder Committee sectors and designated representatives

SECTOR	**Representative**
Local Government	John Wheeler (SWRCB)
Federal Government	Debra Denton (EPA Region IX)
Regional Water Board; NPDES Permitting	Veronica Cuevas (RWQCB4)
Wastewater Agencies	Mitch Mysliwiec (LWA representing CASA)
Stormwater Agencies	Jian Peng (CASQA)
Agricultural Coalition	Sarah Lopez (Central Coast, CCWQP)
Non-Governmental Organization	Kaitlyn Kalua (CA Coastkeeper)
Private Laboratories	Jeff Miller (Aqua-Science Laboratories)
Public Laboratories	Josh Westfall (Los Angeles County Sanitation Districts)

Table 2. List of Expert Science Panel disciplines and designated scientists.

Expertise	**Representative**
Freshwater Toxicology- Academic	Robert Brent (James Madison University)
Freshwater Toxicology- Government	Teresa Norberg-King (USEPA)
Freshwater Toxicology- Industry	Howard Bailey (Nautilus Environmental)
Biostatistics	A. John Bailer (Miami University)
Data Quality Objectives for WET testing	Leana Van der Vliet (Environment Canada)

Inventory of lab techniques and analyses of historical data

Inventory of Laboratory techniques

To date, no one has created an inventory of laboratory techniques used by ELAP accredited laboratories for the *C. dubia* reproduction test. This task will create that inventory, some of which will be easily accessible and expected based on the promulgated method, and some of which is expected to be more difficult. The inventory will focus on four types of lab techniques including:

- Dilution water
- Food
- Culturing
- Technician training and laboratory's level of experience

Table 3 distills the factors within each of the four types of data to be inventoried, and how they will be collected. For the most part, these categories correspond to parts of the promulgated method that allows for flexibility in their lab techniques.

Multiple approaches will be used to compile the lab technique information including Standard Operating Procedures (SOPs), supporting documents such as bench sheets and quality assurance plans, and a survey questionnaire. The questionnaire will be created after reviewing SOPs, ensuring that all of the relevant information can be collected. If necessary, follow-up one-on-one interviews with lab managers or lab directors may be conducted to verify lab techniques and fill in any missing information.

Compile historical test data

Historical testing data will be compiled from ELAP accredited laboratories. Testing data will focus on controls and each laboratory's reference toxicant results. Test acceptability requirements for controls dictate minimum lab performance and, because there is a lack of toxicant exposure, it is assumed that this represents each lab performing the test to the best of their ability. Similarly, reference toxicant test requirements dictate specified lab performance, especially in precision of test organism response. Test sample data response will not be utilized because test performance expectations are not known.

The goal is to compile the daily number of neonates per replicate for each test and supporting data. Table 3 lists the factors to be calculated from compiling this historical test data including:

- Average number of neonates/female
- Average number of broods/female
- Age of females at test start
- Control variability as standard deviation (SD), coefficient of variation (CV)
- Reference toxicant 50% lethal concentration (LC50) variability as SD and CV
- Reference toxicant 25% inhibitory concentration (IC25 for reproduction) variability as SD and CV
- Test water quality data
- Number of replicates tested

In addition, laboratories may be queried for testing details including:

- Procedure for determining mortality
- Procedure to exclude 4th broods
- Frequency of test failures (if data not provided)

Table 3. Categories of information types to be collected and their likely method of collection.

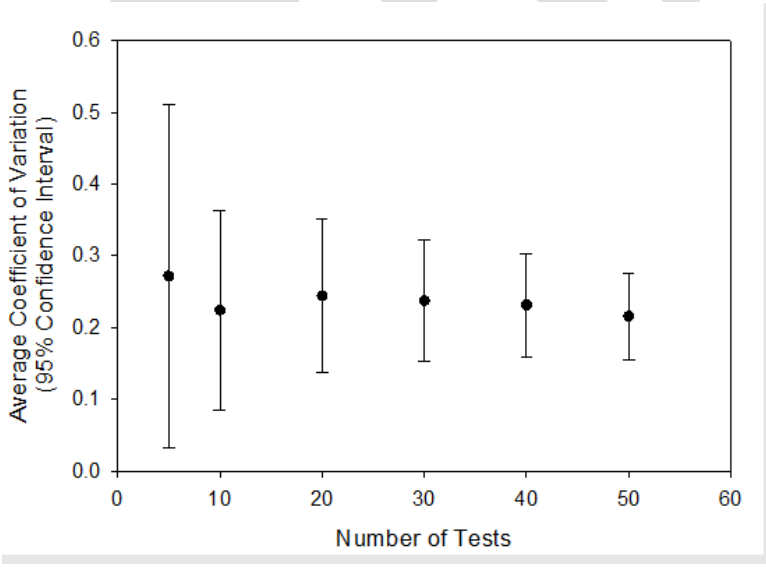
	SOP and supporting documents (e.g. QAP, bench sheets)	Survey and/or phone call	CETIS report and raw data
Dilution water			
Recipe (incl. supplements), vendor	x	x	
Source water	x		
Shelf-time	x	x	
Food			
YCT recipe , vendor	x	x	
Shelf-time	x		
Algal species, source, culture media	x	x	
C. dubia culture			
Frequency of restart/turnover		x	
Frequency of culture failure		x	
Photoperiod	x		
Culture water quality data (e.g. hardness)	x		
Testing procedure/ historical data			
Control variability			x
Age of females at test initiation	x		x
# of neonates/female in controls			x
# of broods/female in controls			x
Daily number of neonates			x
Reference toxicant variability			x
Frequency of test failures	x		x
Test water quality data		x	x
Number of replicates			
Treatment of data outlier		x	
Procedure to exclude 4th broods	x	x	
Time to reproduction	x		
Technician Experience			
Training protocols		x	
Technical experience	x		

Sample size for the historical data compilation is a critical study element. Currently, there are 17 ELAP accredited laboratories conducting the *C. dubia* test in California. Thus, sample size to compare differences among labs is already truncated. The study goal is to collect data from every ELAP accredited laboratory.

Sample size for within laboratory variability assessment is also critical. The study goal is to compile at least 30 tests or the last 3 years of test data from every ELAP accredited laboratory, whichever comes first. Knowing that laboratories do not have the same testing frequency, this target sample size is based on weighing two competing factors that could influence variability assessments. The first factor is the desire to have as much data as possible to have confidence in the variability quantification. The second factor is the desire to keep data as current as possible to reduce the effects of potentially variable-inducing parameters such as evolving lab techniques, turnover in personnel, and other challenges.

Figure 2 presents the results of a simulation to assess confidence in control variability at various sample sizes. This simulation focused on the coefficient of variation (CV) in control brood size per female. Lower CVs have less variability than larger CVs, and Fox et al (2019) targets CVs<0.25 as preferred for reducing errors using the TST. For this simulation, the number of neonates per female ranged from 15 – 30, which meets the promulgated method minimum and approximates the example used in USEPA (2012). The number of neonates per female was randomly assigned for 10 replicates per test and the control CV calculated. This was repeated for sample sizes ranging from 5 to 50 tests. The average and 95% confidence interval of brood size CV per test was calculated for varying sample sizes in Figure 2. This simulation illustrates two important points; a) the average CV tends to stabilize after about 20 tests, and b) the 95% confidence interval about the average CV continues to get smaller with more tests (as expected), but large gains in confidence subside after sample sizes >30. These two results support the study goal of 30 tests or 3 years, whichever comes first.

Figure 2. Simulation to assess average and 95% confidence intervals in control coefficient of variation (CV) at various sample sizes.



Data management, quality assurance and analysis

Data workflow will follow the schema in Figure 3. Input data will come in two categories mirroring the subtasks for this step: laboratory techniques and historical data.

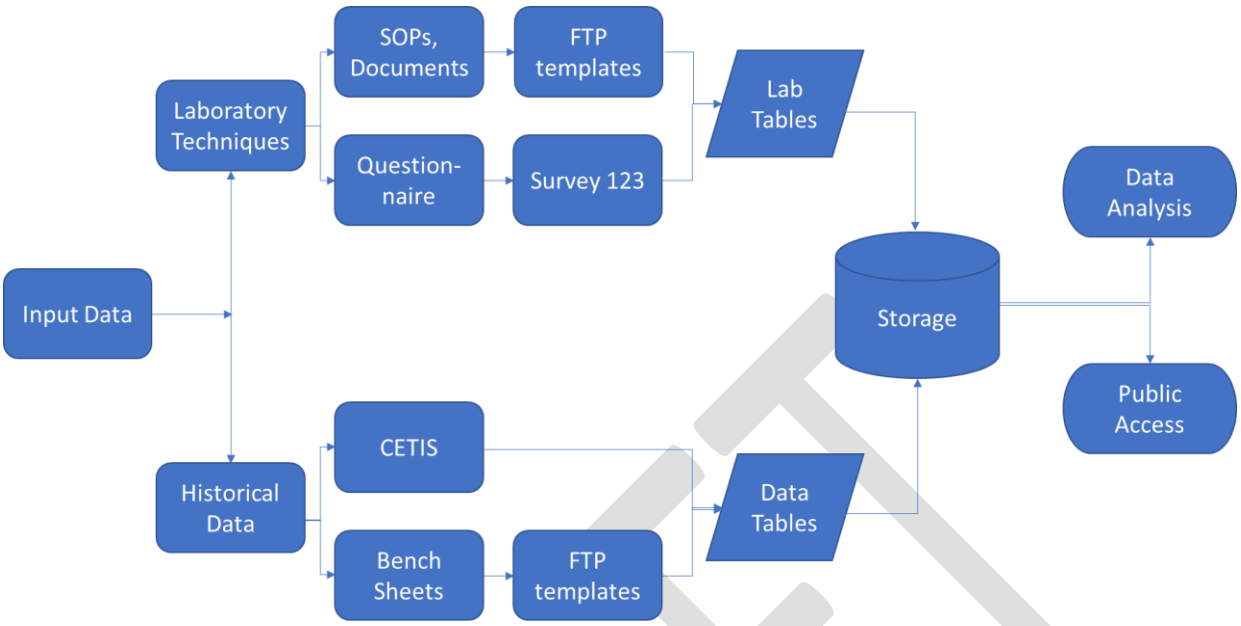
Laboratory technique data will be gathered in two different formats. The first format will be document extraction. This includes documents such as SOPs, Quality Assurance Plans, and other supporting documentation. These data will be categorical, and input using file transfer protocol (ftp) data entry templates by SCCWRP staff. At least 20% of the hand data entered will be checked manually by supervisors and, if errors are found, the remaining 80% of data from that data entry personnel will be checked and accompanied by additional personnel training. The ftp process will also include automated data checkers prior to storage for a 100% data audit of completeness, redundancy, syntax, and look up list errors. The second format will be the laboratory survey, where directed questions will be asked of lab managers or lab directors. The electronic survey will be created in Survey 123, with pre-defined look up lists to eliminate any data entry errors. After automated data checks for completeness and redundancy, these data will be loaded into storage.

Historical data will be more data management intensive than laboratory technique data. The preferred route for compiling historical data is to utilize CETIS (Comprehensive Environmental Toxicology Information System), a software package utilized by regulated agencies for submitting compliance toxicology data. CETIS provides all of the data necessary for this project including daily counts of surviving females and neonate production per replicate, as well as water quality monitoring. SCCWRP will create an ftp for receiving exported CETIS files and transforming into the necessary formats for this project. Automated data checkers prior to storage will create a 100% data audit of completeness, redundancy, syntax, and look up list errors. Where errors occur in exported CETIS data, SCCWRP will query the laboratory of origin for the missing data.

For historical data that is not in CETIS, SCCWRP will need to hand enter data directly from bench sheets. Similar to the workflow for laboratory techniques, these continuous data will utilize file transfer protocol (ftp) data entry templates input by SCCWRP staff. At least 20% of the hand data entered will be checked manually by supervisors and, if errors are found, the remaining 80% of data from that data entry personnel will be checked and accompanied by additional personnel training. The ftp process will also include automated data checkers prior to storage for a 100% data audit of completeness, redundancy, range checks, syntax, and look up list errors.

Once completed, the central data storage will link the laboratory technique and historical data through unique key fields to bind the tests specific to each laboratory. All laboratories will remain anonymous for this study to enhance laboratory participation. The central data storage will be located behind firewalls on SCCWRP servers and backed up twice daily in at least two locations to ensure data security. The laboratory technique and historical data will become publicly accessible data at the conclusion of the study. However, all laboratory identifiers will be kept anonymous to ensure that laboratories are free to participate without fear of being singled out.

Figure 3. Data workflow for this study.



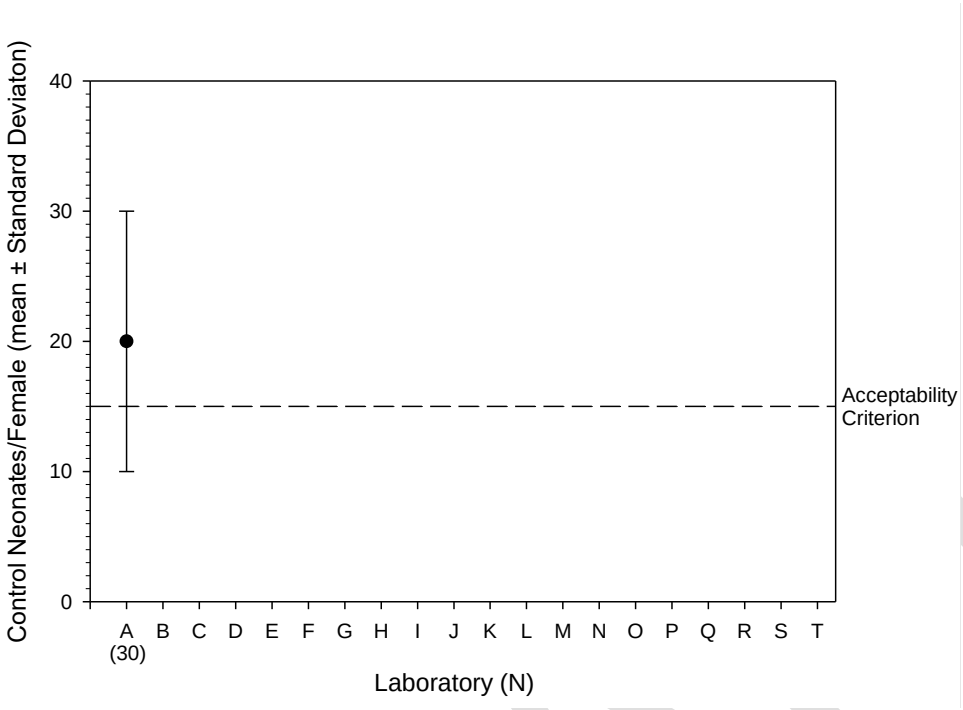
Data analysis will focus on three aspects for this task, consistent with the study questions:

1. Lab technique inventory
2. Comparison of lab performance among laboratories based on historical data
3. Relationships between lab performance and historical data

The lab technique inventory will be a table that quantifies the number of labs utilizing each technique.

Comparison of lab performance among laboratories will utilize pairwise or multiple pairwise (e.g., ANOVA) analysis between laboratories for key test performance metrics. Key test performance metrics will focus on the number of neonates per female (average, standard deviation, CV) in controls and reference toxicant effect concentrations (LC50 and IC25, average, standard deviation, CV). An example of one keystone graphic to compare lab performance among laboratories is Figure 4. To be clear, while statistical analysis will be applied to these data, statistically significant differences are not the goal of this analysis. Especially with so few labs and expected unequal sample sizes among labs, it is anticipated that patterns will be just as important as statistical testing. Where patterns emerge, the lab techniques for the best and worst performing laboratories can be more carefully examined.

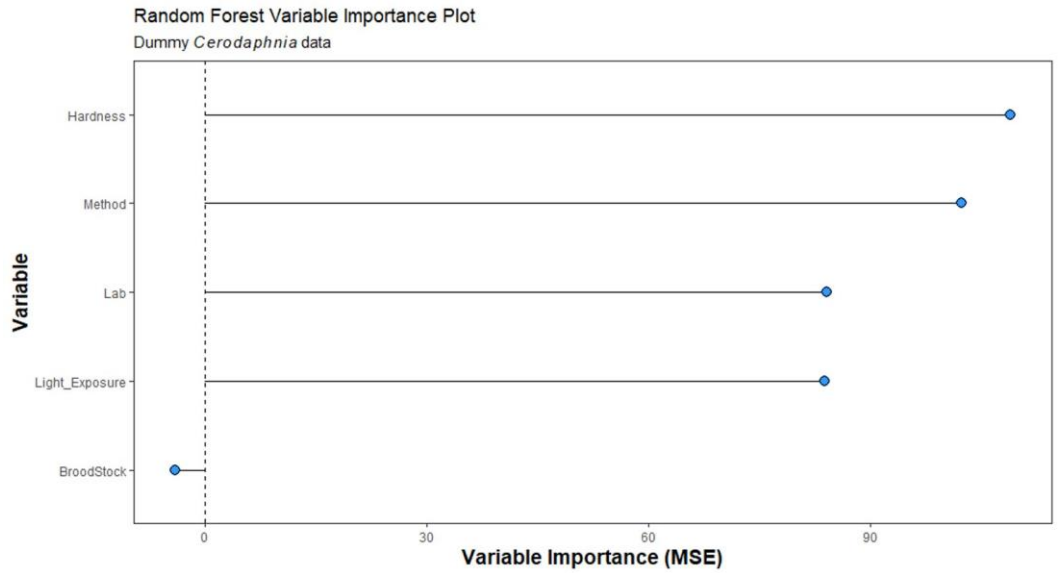
Figure 4. Example keystone graphic for comparing control performance among laboratories. Data for Lab A is not real and plotted only as an example. Data is not plotted for Labs B-T.



Relationships between lab techniques and historical data is the ultimate goal of this task. Multivariate statistics will be used to compare categories and subcategories of lab techniques and the historical data test metrics to determine which lab technique(s) is the source of the most variability (e.g., dilution water recipe versus food source, versus laboratory experience, etc.). The multivariate analysis will include Random Forest and/or Linear Mixed Effects Models, which will attempt to identify variables of greatest importance. An example keystone graphic may look like Figure 5, where mean square error is used to rank variable importance. The data in figure 5 are not real and should be used only for visualization of example end product.

From this series of analyses, a set of proposed revisions or standardizations to the laboratory techniques will be developed. If it is determined that it would be useful to gather more data from the laboratories to verify what was established during the data analysis, then round-robin testing may be proposed if there is sufficient funding. This exercise would seek the participation of all ELAP accredited laboratories to analyze select split samples (e.g., dilution water with different hardness).

Figure 5. Example of keystone graphic illustrating results of random forest. Data is not real, and is for illustration purposes only.



~~Dose-response~~ testing to optimize lab techniques

The third of five steps in this study will focus on ~~dose-response~~ testing procedures to quantify the variability of lab techniques identified by the historical analysis in Step 2.

There may potentially be many variable inducing differences in lab techniques from Step 2, and not all of them may be followed up in this Step 3. So, a prioritization of which lab techniques require ~~dose-response~~ testing will need to be evaluated by the Stakeholder Committee and Expert Science Panel formed in Step 1. The criteria for prioritization should include:

- Lab techniques that appear to contribute the greatest within or among test variability
- Lab techniques that have multiple options across the most laboratories
- Techniques that have optional approaches in the promulgated method guidance
- Techniques that are not defined in the promulgated method guidance
- Others as agreed upon by the Stakeholder Committee

The ~~dose-response~~ testing will be conducted by one (or a limited number) of laboratories to remove confounding interlaboratory variability. This should help to isolate the variability associated with only the lab technique. Therefore, the laboratory(ies) selected should be amongst the most competent and experienced to ensure capability, minimize intra-laboratory variability, and need not be ELAP accredited.

The results of the ~~dose-response~~ testing will culminate in a technical memorandum that describes the variations of the lab technique selected to be tested, the results of the ~~dose-response~~ testing, and a draft recommended guidance on the optimized lab technique to produce the least variability in test results.

Evaluation of the revised lab technique

The fourth of five steps in this study will be for all of the ELAP accredited laboratories to participate in a round-robin split-sample exercise using the draft laboratory technique guidance developed from Step 3.

Since the project builds from step-to-step, the exact sample types, number of laboratories, and the laboratory technique guidance are currently unknown. The sample types will be chosen so that the effect of the draft laboratory technique guidance can best be tested and quantified. It is presumed that the challenge samples may include a variety of blank samples, as well as some spiked samples. However, the exact number of samples will be agreed upon after Step 3. The number of laboratories will be dependent upon the amount of additional resources available.

Whatever criteria are used to select samples for testing, they first will be vetted by the Stakeholder Committee and then approved by the Expert Science Panel.

The results of the split-sample testing will culminate in a technical memorandum that describes the split-samples created and distributed to laboratories, the results of the split-sample testing, and an assessment of the final draft recommended guidance on the optimized lab technique to produce the least variability in test results.

Final report with final recommended guidance

The final report will summarize the study objectives, methods, results, and a discussion of the findings and limitations of the study. The final report will include the interim deliverables contained within the Technical Memos from Steps 2-4. The final report will also be published documentation to accompany the project database.

Most importantly, the final report will contain the vetted recommended guidance for laboratory activities to optimize variability implementing the *C. dubia* reproduction test.

The Stakeholder Committee will have multiple opportunities to review and provide input on the final report. The Expert Science Panel will also review the final report and provide a consensus opinion on the recommended laboratory technique guidance for implementing the *C. dubia* reproduction test.

SWRCB staff will be responsible for deciding the final disposition of the recommended laboratory technique guidance, and the final recommendation to the State Water Board.

Schedule

Task	Product	Deadline
Study Workplan		
Draft	Draft Workplan to identify potentially variable-inducing lab techniques	3/1/21
Final	Final Workplan approved by Expert Science Panel	5/1/21
Historical Data Analysis		
Lab Data Analysis	Technical Memo identifying potentially variable-inducing lab techniques	7/1/21
Split Sample Analysis (if conducted)	Technical Memo quantifying within and among lab variability	1/1/22
Optimization Testing	Technical memo with draft recommended guidance to reduce within and among lab variability	3/30/22
Interlaboratory Testing	Technical Memo quantifying within and among lab variability	7/31/22
Final Report		
Draft	Draft Report with final recommended guidance	11/1/22
Final	Final Report approved by Expert Science Panel	12/31/22

References

Diamond, J., P. Stribling, M. Bowersox, and H. Latimer. 2008. Evaluation of effluent toxicity as an indicator of aquatic life condition in effluent dominated streams: A pilot study. *Integrated Environmental Assessment and Management* 4:456-470.

Fox, J.F., D.L. Denton, J. Diamond, and R. Stuber. 2019. Comparison of false-positive rates of 2 hypothesis-test approaches in relation to laboratory toxicity test performance. *Environmental Toxiciology and Chemistry* 38:511-523.

Moore, T.F., S.P. Canton, and M. Grimes. 2000. Investigating the incidence of type I errors for chronic whole effluent toxcitiy testing using *Ceriodaphnia dubia*. *Environmental Toxiciology and Chemistry* 19:118-122.

Schiff, K.C. and D. Greenstein. 2016. Stormwater monitoring coalition toxcity testing laboratory guidance document. Southern California Coastal Water Research Project. Costa Mesa, CA.

USEPA. 2001. Final report: Interlaboratory variability study of EPA short-term chronic and acute whole effluent toxicity test methods, Vol. 1. U.S. Environmental Protection Agency. Washington, DC.

USEPA. 2002. Short-term methods for estimating the chronic toxicity of effluents and receiving water to freshwater organisms. EPA-821-R02-013. U.S. Envirnmental Protection Agency. Washington DC.

APPENDIX A – List of ELAP Accredited Laboratories

Lab Name	**Lab Type**
ELAP accredited laboratories in California	
49er Water Laboratory	Private
Aqua-Science	Private
Aquatic Bioassay & Consulting Laboratories, Inc.	Private
Aquatic Testing Laboratories	Private
Aquatic Toxicology Laboratory, Aquatic Health Program	Academic
Enthalpy Analytical, LLC (Nautilus)	Private
Environmental Monitoring Div. (EMD) Lab. at Hyperion Treatment Plant	Public
Granite Canyon -- UC Davis Marine Pollution Studies Laboratory	Academic
Inland Empire Utilities Agency Laboratory	Public
MBC Aquatic Sciences	Private
McCampbell Analytical, Inc.	Private
Pacific EcoRisk	Private
San Jose Creek Water Quality Laboratory	Public
Wood Environment & Infrastructure Solutions, Inc.	Private
ELAP accredited laboratories outside of California	
Eurofins TestAmerica - Corvallis (ASL)	Private
GEI Consultants, Inc.	Private
Ramboll	Private
Tetra Tech's Ecological Testing Facility	Private

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
REGION IX
75 Hawthorne Street
San Francisco, CA 94105

Thank you for the opportunity to review and comment on the March 18 SCCWRP draft "Conceptual Workplan: Quality Assurance Evaluation of the *Ceriodaphnia dubia* Reproduction Test." These are my comments:

Overall Comments

1. The workplan lacks a detailed schedule including very importantly – the development of the Project's Quality Assurance Program Plan (QAPP) --- see EPA for elements of the QAPP: https://www.epa.gov/sites/production/files/2015-06/documents/g5-final.pdf. This draft workplan lacks a project/task organization – very critical for this work is the identification of the project's statistician, QA officer, etc. for this work. The QAPP should include study specifics on project/task organization – list of individuals involved in the study, their roles, and responsibilities, the individual responsible for maintaining the QAPP, etc. Ideally the workplan should follow the data quality objectives process. For example, in the execution of Step #2 – evaluation of historical data – the workplan lacks a description of performance and acceptance criteria expressed in terms of data quality indicators (DQI) such as precision, bias, accuracy, representativeness, comparability, completeness and sensitivity as these apply to each of the workplan steps. Recognizing that some DQI may not be applicable for each workplan step. A detailed data management plan is essential. It should address data to be produced, standards for data/metadata format and content, policies for access and sharing, and archiving and preservation of data access. **The benchmark for this Study is peer reviewed publications with results that are of known quality and publicly accessible**.
2. The historical review should summarize percentiles for control mean, control standard deviation and control CV from each accredited lab. The Science Panel can comment on whether categorizing labs by these metrics will help identify practices/procedures that may cause differences in lab performance. Then, the study evaluation could bin labs long-run cCV as either being above or below value; then with closer examination (say of other data metrics; SOPs, and lab survey responses) to identify potential factor(s) that maybe influencing a given lab's performance.
3. Now to discuss the separate 4-page CASA proposal, received March 22. I am gravely concerned that the CASA proposal is changing the State Board's charge to SCCWRP to understand and evaluate "how to improve lab performance". Any party that provides additional funds for the SCCWRP work must follow the governance rules as defined (input from all stakeholders and the science panel). More importantly, the scientific process must be followed, which is the knowledge obtained from the Historical review (step 2) of all CA accredited labs will lead to what is/are the testable hypotheses to be evaluated, if and when any samples are sent to testing labs in CA for a Round Robin (RR) Study. It appears from the proposal that they are directing the scientific questions to be evaluated in a RR **element without even conducting** the other elements (review Historical data). More importantly --- how this is not prescribing the outcome of the study nor a conflict of interest directly by the entities being regulated? Our Agency understanding was that any additional funds to conduct a RR could be accepted and placed in the State's funding vehicle **BUT, cannot pre-dispose how the funds are to be spend on specific objectives; nor change the clear direction agreed** and stated by SB. Therefore, I will not use my time to provide comments on a premature and in adequate draft proposal at this time.

   Lastly, I am enclosing my detailed comments (see below), comments on the SCCWRP workplan pdf (separate file), and enclosing my comments on previous SCCWRP 2016 Study as often cited as the genesis for the questions raised by the regulated community. If you have any questions, I can be reached via email at denton.debra@epa.gov.

Sincerely,

Debra L Denton, PhD
Environmental Scientist
USEPA Region 9, Water Division

Here are EPA detailed comments and attached are additional detailed comments in the draft March 18 SCCWRP document.

For ease of discussion purposes, I have detailed the Steps of the draft March 18, SCCWRP workplan.

Step	Task	Timeline
1	Create Governance Structure	Done
2	Analysis of historical data	July 1, 2021
3	Dose-response testing	March 3, 2022
4	Evaluation of revised lab technique via split sample testing	July 31, 2022
5	Final Report	December 31, 2022

For ease of some discussion, I am pasting EPA's terminology for defining test variability – as there are several components of test variability. Copied from U.S. Environmental Protection Agency. 2000. Understanding and accounting for method variability in whole effluent toxicity applications under the National Pollutant Discharge Elimination System Program. Eds: Denton DL, Fox J, Fulk FA, Greenwald K, Narvaez M, Norberg-King TJ, Phillips L. EPA/833/R-00-003. Office of Water. Washington, DC. Additionally, see Appendix B which contains technical and explanatory notes, and supplemental tables pertaining to the statistical analyses of reference toxicant test results presented in Chapters 3 and 5.

2.2 Defining WET Test Variability

As with any measurement process, WET tests have a degree of variability associated with the test method performance. Three measures of variability related to WET tests are within-test variability, within-laboratory variability, and between-laboratory variability.

- **Within-test (intra-test) variability** is the variability in test organism response within a concentration averaged across all concentrations of the test material in a single test.

- **Within-laboratory (intra-laboratory) variability** is the variability that is measured when tests are conducted using specific methods under reasonably constant conditions in the same laboratory. Within-laboratory variability, as used in this document, includes within-test variability. The American Society for Testing and Materials (ASTM) uses the term "repeatability" to describe within-laboratory variability. Repeatability is estimated (as a sample variance or standard deviation) by repeating a test method under realistically constant conditions within a single laboratory.

- **Between-laboratory (inter-laboratory) variability** is the variability between laboratories. It is measured by obtaining results from different laboratories using the same test method and the same

test material (e.g., reference toxicant). Between-laboratory variability, as used in this document, does *not* include the within-laboratory component of variance. ASTM uses the term "reproducibility" to describe between-laboratory variability. Reproducibility is estimated by having nearly identical test samples (duplicates or splits) analyzed by multiple laboratories using similar standard methods. Although reproducibility is generally synonymous with between-laboratory variability, estimates of reproducibility may combine within-laboratory and between-laboratory components of variance, making between-laboratory variability numerically larger than within-laboratory variability as defined above.

For purposes of consistency, EPA uses the terms within-laboratory and between-laboratory variability throughout this document.

Numerous factors can affect the variability of any toxicity test method. These factors include the number of test organisms, the number of treatment replicates, randomization techniques, the source and health of the test organisms, the type of food used, laboratory environmental conditions, and dilution water quality. The experience of the analyst performing the test, analyzing the data, and interpreting the results may also affect variability (Grothe et al. 1996, Fulk 1996).

2.3 Quantifying WET Test Variability

Review of Step 2: page 6, Compile Historical Data:

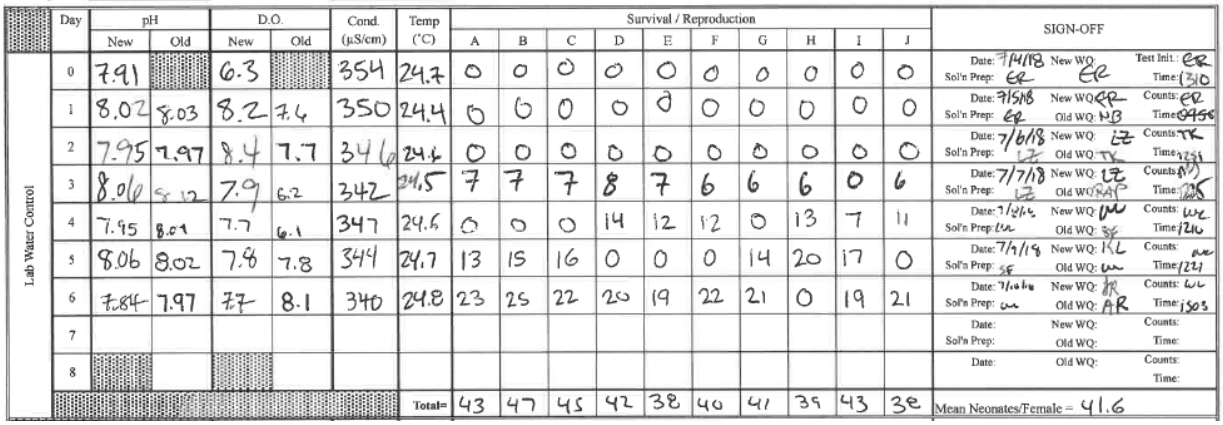
This study should rely on performance criteria (also referred to as measurement quality objectives) that are defined at the outset of the study to evaluate comparability of CA accredited labs. In this case, criteria would be based on various quality control toxicity metrics such as control means, standard deviations, and coefficient of variation (CV) results from both reference toxicant (RT) data and control data generated from effluent tests. Format of basic data parameters for this analysis; each lab provides the last 30 to 50 valid control information to quantify each lab's percentiles for control mean, control standard deviation, and control (cCV). The QAPP must identify the format and data elements for each study objective – briefly here is an example of some data elements to be captured for quantification of control data to evaluate each lab's long-run cCV.

Test Date	# Reps	Control mean	Control Std	Control CV	Water Hardness
2/6/19	10	26.3	6.9	26.40	MH

The QAPP should demonstrate that proposed data analysis using proposed sample sizes will result in finding factors or practices (if they exist) that can lower control CVs (cCV) or increase control means by specified amounts (not relative amounts but absolute differences, for example an increase of 10 neonates or a decrease from CV 0.25 to 0.15). Step 6 of the data quality objectives process (specify performance or acceptance criteria) is lacking in this workplan. How will proposed changes in the test method will be evaluated with statistical confidence that the results are meaningful?

The sample sizes (numbers of labs, number of data for each lab) should support the specific inferences needed in this Study. The quantitative inferences should be identified as objectives. For example: - to show that categorical factor A is associated with a 10% increase in the control average neonate count or 10% decrease on control CV - to identify 2-3 categorical factors that together account for a 10% increase in the control average neonate count or 10% decrease on control CV. The design should be capable of identifying such factors with a specified confidence level if they exist.

It is important to understand that the meta data – at the daily count level is not captured in CETIS reports. Here is a quick snapshot of capturing data at the daily level based on what is expressed in Table 3 of workplan. Point is that a detailed QAPP will define what data elements, and the format is desired prior to execution of the workplan.



	Day	pH New	pH Old	D.O. New	D.O. Old	Cond. (µS/cm)	Temp (°C)	Survival / Reproduction A	B	C	D	E	F	G	H	I	J	SIGN-OFF
Lab Water Control	0	7.91		6.3		354	24.7	0	0	0	0	0	0	0	0	0	0	Date: 7/4/18 New WQ: ER Test Init.: ER Sol'n Prep: ER Old WQ: Time: 1310
	1	8.02	8.03	8.2	7.4	350	24.4	0	0	0	0	0	0	0	0	0	0	Date: 7/5/18 New WQ: ER Counts: ER Sol'n Prep: ER Old WQ: NB Time: 0945
	2	7.95	7.97	8.4	7.7	346	24.6	0	0	0	0	0	0	0	0	0	0	Date: 7/6/18 New WQ: LZ Counts: TK Sol'n Prep: LZ Old WQ: TK Time: 1254
	3	8.01	8.02	7.9	6.2	342	24.5	7	7	7	8	7	6	6	6	0	6	Date: 7/7/18 New WQ: LZ Counts: NV Sol'n Prep: LZ Old WQ: RAP Time: 1225
	4	7.95	8.01	7.7	6.1	347	24.6	0	0	0	14	12	12	0	13	7	11	Date: 7/8/18 New WQ: NV Counts: WL Sol'n Prep: LZ Old WQ: JS Time: 1216
	5	8.06	8.02	7.8	7.8	344	24.7	13	15	16	0	0	0	14	20	17	0	Date: 7/9/18 New WQ: KL Counts: AR Sol'n Prep: JS Old WQ: WL Time: 1221
	6	7.84	7.97	7.7	8.1	340	24.8	23	25	22	20	19	22	21	0	19	21	Date: 7/10/18 New WQ: AR Counts: WL Sol'n Prep: WL Old WQ: AR Time: 1303
	7																	Date: New WQ: Counts: Sol'n Prep: Old WQ: Time:
	8																	Date: Old WQ: Counts: Time:
							Total=	43	47	45	42	38	40	41	39	43	38	Mean Neonates/Female = 41.6

Table 3, page 7:

"Frequency of test failures" should be defined more carefully. There could be multiple reasons why a test could be invalidated or stopped. All such events should be counted and distinguished. Of note: EPA Region 9 conducted a review of a major permittee (who does their own toxicity testing) in CA. EPA found that the lab had a pattern of declaring valid effluent tests as invalid based on the reference toxicant data results, without providing the rationale for invalidating effluent test results. These details are important in reviewing lab performance. The current study needs to obtain and review the lab's "correction actions" on the invalid tests and how the issue(s) were resolved.

This is **a portion of the review by EPA Region 9** – which highlights some of the data interpretation issues for an accredited permittee based-lab in CA. In many datasets, the laboratory misinterprets USEPA guidance on data evaluation.

A. Failed QA/QC: Numerous toxicity test results were not reported and did not provide proper justification, stating that they "Failed to meet internal QA/QC requirements." This statement lacks detail and reasons for terminating tests should be elaborated in the reports and the QA Plan. Toxicity tests conducted during the entire xxxxx resulted in numerous QA/QC failures despite retesting. No valid data were available for the month of chronic tests with *Pimephales promelas* on effluents of xxxx and chronic tests with *Ceriodaphnia dubia* on effluents from xxxx and xxxx. Therefore, compliance could not be determined.

Page 8, last paragraph "Figure 2 presents...The following are comments by Dr. John Fox:

Some points about this paragraph:

1. Paragraph does not say how number of neonates was simulated (which probability distribution) and how many simulations were run for each number of tests (5-50 tests).
2. This sentence displays poor understanding of statistical principles: "*Lower CVs have less variability than larger CVs ...*" because it does not distinguish parameters from sample

estimates. More accurate revision: "Smaller parametric CVs result in less variable sample CVs."

3. "*The number of neonates per female was randomly assigned for ...*" See #1 - how?
4. "*95% confidence interval*" What method was used to get this confidence interval?
5. "*... but large gains in confidence subside after sample sizes >30. These two results support the study goal of 30 tests or 3 years, whichever comes first.*" First, it is not "large gains in confidence" rather it is width of the confidence interval. Second, for most CI-methods, the confidence interval width will decrease gradually in proportion to the inverse of the square root of number of tests (assuming number of replicates per test remains the same).
6. Still referring to the last sentence of the paragraph: responsible planning would identify a particular value for confidence interval width or standard error of the mean CV that will satisfy project objectives.

Comments on data analysis discussed on pages 10-12…..the following are comments by Dr. John Fox:

Using Figure 2 on page 6 as an example, identify a value of parametric CV that should be distinguishable statistically with 95% confidence and which has practical consequences. For example, should a difference between labs in average CV parameter of 0.10 be distinguished? Set a similar target for control means.

These target values for sample sizes will determine how well the data analysis can identify lab practices that may lead to reduced variability.

Further analysis is needed to show that these sample sizes will allow better practices to be identified reliably. This can be demonstrated using artificial or simulated data.

"*Comparison of lab performance among laboratories will utilize pairwise or multiple pairwise (e.g., ANOVA) analysis between laboratories for key test performance metrics.*" (page 10).
Comment: Some metrics may not satisfy assumptions for inference using standard ANOVA and may require nonparametric statistical methods.

"*Especially with so few labs and expected unequal sample sizes among labs, it is anticipated that patterns will be just as important as statistical testing. Where patterns emerge, the lab techniques for the best and worst performing laboratories can be more carefully examined*."
Comment: What this says is "we think we can identify meaningful patterns, even if they are not significant" If this statement is meaningful, then back it up with a demonstration using simulated or artificial data, showing that the proposed analysis can distinguish better practices.

Review of Step 3: "Dose-response testing" on page 13

Explain how dose-response testing will quantify variability accurately and precisely using only a small number of labs. "*The dose-response testing will be conducted by one (or a limited number) of laboratories to remove confounding interlaboratory variability. This should help to isolate the variability associated with only the lab technique.*" The nature of the split samples is not described. Does "dose-response testing" mean that reference materials will be used – if so, what is the SOP of sample execution? **This step of the workplan lacks necessary details.**

Comments about planned round-robin study on page 13:

This component of the workplan is so vaguely described in the SCCWRP workplan that detailed comments cannot be expressed. A RR study – would need a detailed workplan with testable study hypotheses, describe the experimental design including detailed sample size with consultation with statisticians, a QAPP, of which must be reviewed by both the Stakeholder and Science Panels. Without proper statistical input --- the study will not be able to distinguish differences observed among labs for the testable hypotheses. This study component ---- must have detailed SOPs on how samples will be formulated, confirmed, and executed by an independent source, along with having a referee testing laboratory, in the execution of any RR.

Since, the CASA draft is re-introducing the SCCRWP 2016 Study report; therefore, I am enclosing the EPA comments on the draft SCCWRP 2016 study, which was used to create uncertainty about the reproduction endpoint and has been part of the genesis of this current Study **(enclosure below).**

Enclosure: Denton and Diamond comments on the draft 2016 SCCWRP report.

Background

On July 1, 2016, Ken Schiff of the Southern California Coastal Water Research Project (SCCWRP) emailed Debra Denton of EPA Region 9 drafts on a "Stormwater Monitoring Coalition Toxicity Testing Laboratory Guidance Document," dated June 2016, and "Round 2: Stormwater Toxicity Testing Laboratory Intercalibration Data Analysis Standard Operating Procedures," dated December 2015. Dr. Denton and EPA consultant Dr. Jerry Diamond of Tetra Tech reviewed the documents and provided the comments below.

Comments on "Stormwater Monitoring Coalition Toxicity Testing Laboratory Guidance Document"

1) The approach used to evaluate interlaboratory comparability is incorrect and misleading. Dr. Diamond served as the lead technical contractor for the National Methods and Data Comparability Board, which is a workgroup under the National Water Quality Monitoring Council. The Board published materials describing ways to evaluate comparability of different labs for method-dependent parameters, which includes toxicity. According to this inter-agency workgroup and the Council, a study should rely on method performance criteria (also referred to as measurement quality objectives) that are defined at the outset of the study in order to evaluate comparability of labs for a method-dependent parameter. In this case, criteria would be based on various quality control toxicity metrics such as control coefficient of variation (CV) within tests, CVs for IC25s for *Ceriodaphnia* reproduction in reference toxicant tests, percent minimum significant difference (PMSD) values in reference toxicant tests, and range of 95% confidence intervals for point estimates (i.e., EC25 in the case of this study) in the copper exposure tests. An example using these types of metrics and associated performance criteria can be found in Table 3 titled "Measurement quality objectives for WET tests" of the peer-reviewed paper by Diamond J, Stribling J, Bowersox M, Latimer H. 2008. *Integrated environmental assessment and management*. 4(4) 456-470. See attached file with Dr. Diamond's specific comments within the SCCWRP draft.

2) The reviewers did not receive the actual lab data and thus cannot assess whether the results presented are accurate and whether there may have been other technical issues with the study.

3) It is not clear whether the document should be called "guidance" because its source, SCCWRP, is not a regulatory agency.

4) It is unclear whether there were nine or 10 laboratories participating in the study.

5) The study design was not created properly to evaluate and understand intra- and interlaboratory test precision. See USEPA. 2001a. Final report: Interlaboratory variability study of EPA short-term chronic and acute whole effluent toxicity test methods, Volume 1. Office of Water. Washington, DC. EPA/821/B-01/004, and USEPA. 2001b; Final report: Interlaboratory variability study of EPA short-term chronic and acute whole effluent toxicity test methods, Volume 2. Office of Water. Washington, DC. EPA/821/B-01/005. The study is set up to provide additional specificity when following the EPA methodology to get more comparability among laboratories in this particular watershed for stormwater testing. If precision had been fully evaluated all the participating laboratories would have provided their ongoing control charts for their reference toxicant for each test species and endpoints, as well as providing the laboratories' control performance as both mean and standard deviation for at least the last 20 tests for each test species and endpoint. This is the information that is needed to assist in evaluating a laboratory's performance over time with the same testing conditions, meaning same test organism supplier, same type and composition dilution water, same feeding regime, same glassware type, etc.

6) Significantly, this Study lacks a chemical confirmation conducted for Round 1. In fact, under Section 3.1 Approach and Methods, where is the documentation (either an attachment to the Study) of how homogenized samples were created and confirmed by SCCWRP? This point is critical. The reviewers could not verify that the samples sent to the labs were truly what the study claims. For instance, the reviewers would have liked to read documentation that demonstrates that the dilution water (LDW) was truly dilution water. Similarly, since information on the EC25 for copper is known for the test species in the study, it would have been beneficial to compare EC25s and corresponding 95% CI between reference toxicant EC25s and EC25s obtained in this study (with hardness adjustment if necessary to account for different hardness in dilution water used) for those labs that use copper routinely for reference toxicant testing.

7) Under Approach and Methods, for item # 4, there is no documentation of how the samples were created and distributed among the laboratories. This documentation is paramount. Significantly, Round 1 samples of the "non-toxic" blanks were created by one of the participating laboratories (see the SCCWRP May 2015 SOP). This renders the results questionable. Specifically, the information under 3.2 for assessment of *C. dubia* is suspect because the Round 1 non-toxic samples were created by a participatory laboratory.

8) The tables 10-15 footnote b = effects concentration at 25th percentile is unclear to the reader. It is unclear how these values were calculated. These comments are applicable to the remaining test species tables of summary mean responses. The summary data table should be reporting the EC25 values and associated 95% confidence intervals. This a major question of how the test results were evaluated. To calculate and evaluate the labs' results, this report should have calculated for each sample the mean and standard deviation to calculate CV values. These CV values should be compared to EPA 2000 or EPA 2010 tables which provide the cumulative frequency of test results among over 40 labs and more than 500 tests for *Ceriodaphnia* reproduction. The reliance on percent effect in 100% sample is not an appropriate metric for comparison, especially for a well- studied reference toxicant such as copper, for which the literature has extensive data regarding EC25s for the test species and copper using the same test methods. Lastly, a question regarding the LDW percent effect – what specific test concentration is being compared to the control?

9) Is the use of very hard water really a non-toxic blank when compared to moderately hard laboratory water? Additionally, for making the copper-spiked laboratory water was very hard or moderately hard water used?

10) Meta data (associated QA/QC information) for this study were not available which would enable a reviewer to review and document such information as the Chain Of Custody's (to verify that the samples were shipped and received by the lab within proper temperature requirements for the method), the water quality measurements during the study, along with the laboratory bench sheets to verify the reported values. When EPA conducted the 2001 study all of these data were available to calculate and verify the reported values.

11) Throughout, the document needs to refer to the *Ceriodaphnia dubia* as a 6-8 day test not a 7 day test.

12) Suggest modifying text in Executive Summary about providing specific guidance to participants. While some additional specificity to the promulgated method for *Ceriodaphnia* testing, for example, was provided in this study, several steps in the EPA method that are left up to the lab were not specified in this study. These steps include organism feeding, sample handling, procedures for performing daily water renewals, and age of organisms initially used. Since this study is striving to further refine/clarify areas in the test methods to provide consistency among labs, such additional specificity seems warranted. For example, the test chamber type (glass versus plastic) should be standardized among labs when working with runoff samples that could have hydrophobic pollutants such as pyrethroids. It is well-documented that glass and improved sampling and water mixing is necessary to obtain accurate toxicity results for pyrethroids. The SMC is encouraged to provide further specificity for toxicity testing for their watershed program to help improve performance and comparability of results among labs.

13) Consider using consist terminology of laboratory techniques following EPA toxicity test methods throughout the report. These are used interchangeably and incorrectly, for example: "Inventories of analytical efforts among regional contract laboratories indicated differences among laboratory methods and this raised concerns amongst SMC member agencies about data comparability." (Executive Summary). Proper terminology should be laboratory techniques rather than laboratory methods. The laboratory should be following specified test methods such as those from EPA. There are laboratory techniques that are spelled out in a laboratory's SOP for more detail for items such as test chamber type, use of a specific reference toxicant which are a choice within the laboratory.

14) Table 6 in reference to *H. azteca* needs to include the 2002 Acute Test Methods Manual USEPA. 2002. Methods for measuring the acute toxicity of effluents and receiving waters to freshwater and marine organisms. Fifth Edition. Office of Water, Washington, DC. EPA/821/R-02/012. Suggest that Table 6, Recommended Test Conditions and TAC, should be reviewed by the State Water Board's SWAMP program for statewide standardization. This is a State specific decision.

15) Under 2.3.4, we support providing additional clarification on sample handling, such as when dealing with hydrophobic pesticides in watersheds. See USGS (2009). Hladik, M.L., Orlando J.L., and Kuivila, K.M., 2009, Collection of pyrethroids in water and sediment matrices: development and validation of a standard operating procedure: U.S. Geological Survey Scientific Investigations Report 2009–5012, 22 p. Available at http://pubs.usgs.gov/sir/2009/5012/. This report helps to provide more test method specificity, which would help generate better comparability among laboratories when using the same handling and sampling techniques. These are examples of providing additional specificity beyond the standardized EPA methods.

16) Under the discussion of precision guidance, include both:

USEPA. 2001a. Final report: Interlaboratory variability study of EPA short-term chronic and acute whole effluent toxicity test methods, Volume 1. Office of Water. Washington, DC. EPA/821/B-01/004.

USEPA. 2001b. Final report: Interlaboratory variability study of EPA short-term chronic and acute whole effluent toxicity test methods, Volume 2. Office of Water. Washington, DC. EPA/821/B-01/005.

In the same paragraph under precision guidance the authors are confusing EPA's interlaboratory study such as EPA 2001a and 2001b with studies/reports that have reviewed and evaluated inter-and intra-test precision such as Warren-Hicks et al., 2000. It is appropriate to include EPA (2000). U.S. Environmental Protection Agency. 2000. Understanding and accounting for method variability in whole effluent toxicity applications under the National Pollutant Discharge Elimination System Program. Eds: Denton DL, Fox J, Fulk FA, Greenwald K, Narvaez M, Norberg-King TJ, Phillips L. EPA/833/R-00-003. Office of Water. Washington, DC.

17) Section 3.3, on page 21, does not describe the potential problems with the LDW composition and documentation in Round 1 which impact the evaluation of the *C. dubia* test results. The inclusion of DeGraeve et al. (1992) which showed that 44% of their inter-calibration tests failed to be initiated due to unsuccessful cultures is extremely outdated information (pre-2002 methods). It would be more appropriate to use the extensive 2001a and 2001b study results.

18) The report should acknowledge the literature regarding how differences in hardness between test dilution water and culture water routinely used by the lab may influence lab performance of the *Ceriodaphnia* reproduction test and resulting comparability among labs.

19) Under Section 4.1 regarding future intercalibration study designs, the recommendation to align the program's test design with the study design is a good point.

Comments on "Round 2: Stormwater Toxicity Testing Laboratory Intercalibration Data Analysis Standard Operating Procedures"

1) If this is the data analysis SOP for Round 2, where is the data analysis SOP for Round 1? This documentation needs to be provided and if it is not consistent with this SOP, then we are comparing apples to oranges when comparing Round 1and Round 2 test results.

2) Under the discussion in the report regarding a lab retest if test acceptability criteria were not met, it is not clear how this was reported or handled in terms of the comparability analysis. Were there any re-tests conducted in this study, particularly for the runoff samples? Holding time could have been an issue in terms of resulting comparability in that case.

3) Can you measure precision with LDW?

4) Section II (3) of the SOP states: "This step will provide further interlaboratory variability information." How is this done? The TST is a statistical analysis of the control and a concentration of concern providing an answer of statistical significance or no statistical significance. How is this analysis used for quantification of method precision? Precision is measured by the CV = standard deviation divided by the mean.

5) The SOP states: "we will evaluate the effectiveness of the TST in identifying toxicity compared to traditional statistical methods." (p.6) This SOP recommends running the TST by sample type, dilution, and test species. The TST is statistical approach that only analyzes one effluent concentration to a control and thus the issue of multiple dilutions is not applicable.

Conceptual Workplan: Quality Assurance Evaluation of the *Ceriodaphnia dubia* Reproduction Test

Prepared by the Southern California Coastal Water Research Project

For the State Water Resources Control Board

Agreement #19-278-0780

Draft - March 18, 2021

Contents

Introduction

The California State Water Board recently adopted Toxicity Provisions, which include numeric effluent limitations for aquatic toxicity. Implementation of the monthly median effluent toxicity limitation for the *Ceriodaphnia dubia* (*C. dubia*) reproduction test has been delayed until January 1, 2024 for some dischargers as specified in the Toxicity Provisions. During this delay, the State Water Board has committed to a study, in collaboration with stakeholders and laboratories, to investigate factors that can lead to test variability. High test variability could lead to low statistical confidence in assessments of toxicity or non-toxicity.

The *C. dubia* test has been used for many decades and the currently promulgated protocol for the *C. dubia* reproduction test was established nearly 20 years ago (USEPA 2002). The promulgated method allows laboratories some flexibility when implementing certain laboratory techniques. For example, there are multiple options for preparing culture water or food. In some instances, the promulgated method is silent on test techniques, leaving laboratories to use their best professional judgement. To be clear, the *C. dubia* method is well-established and validated. However, these small differences between laboratories may lead to intra- or interlaboratory variability, which could influence test results.

Previous studies have assessed the variability of the *C. dubia* test results within and among laboratories. In the early 2000s, an interlaboratory comparison exercise performed by the EPA found that only 22 out 122 *C. dubia* chronic tests did not meet test acceptability criteria for survival or reproduction (USEPA 2001). The invalid tests were confined to 10 out of 34 participating laboratories. The study reported intra- and inter-laboratory coefficients of variation for the IC25 values of effluent and receiving water split samples at 17% and 28%, respectively for the reproduction endpoint. More recently, a smaller intercomparison exercise was conducted in California to evaluate the reliability of *C. dubia* chronic test for stormwater toxicity evaluation (Schiff and Greenstein 2016). Of the nine labs that tested split samples of dilution water, three were considered "low" comparability. Lack of comparability among a minority of laboratories testing split samples of dilution water was also identified by others (Moore et al. 2000; Diamond et al. 2008). Most recently, routine testing data generated by eight California-accredited laboratories was examined by Fox et al. (2019) and results indicated that intra-laboratory variability, particularly in controls, influenced whether test samples would be identified as toxic. Either reducing the between replicate variability or increasing the number of replicates improved lab performance.

The State of California Environmental Laboratory Accreditation Program (ELAP) accredits all laboratories conducting analysis for regulatory compliance purposes, including the *C. dubia* test. Currently, there are 17 ELAP accredited laboratories for conducting the *C. dubia* test in California (Appendix A). Accreditation is based on the demonstration that laboratories are following the testing protocols, properly training their staff, keeping accurate records, and demonstrating they can meet data quality objectives for reference toxicant and performance evaluation test samples. While this process demonstrates that a laboratory capably performs a test, it does not address test variability between laboratories or differences in lab techniques that are allowed by the protocols.

Objective of this study

The objective of this study is to evaluate sources of variability in the *C. dubia* reproduction test conducted by California-accredited laboratories and identify potential laboratory technique guidance and/or recommendations to: (a) improve the consistency of the execution of the C. dubia test method

to achieve improved precision (i.e., as measured by the control coefficient of variation) within each testing laboratory; and (b) improve the consistency and comparability of *C. dubia* test results among testing laboratories.

The study will seek to answer the following questions:

1) What are the *C. dubia* chronic reproduction toxicity test laboratory techniques used by Environmental Laboratory Accreditation Program (ELAP) accredited laboratories in the state of California?
2) How does variability in control reproduction and/or reference toxicant response in the *C. dubia* chronic reproduction toxicity test compare amongst laboratory technique differences used by ELAP accredited laboratories?
3) Does standardizing differences in the *C. dubia* chronic reproduction toxicity test laboratory techniques reduce variability in control reproduction and/or reference toxicant response?

Based on the results of this study, a list of suggested best-practices for the *C. dubia* reproduction test laboratory techniques will be developed.

Just as importantly, this study is not designed to address or quantify false negative or false positive rates for detecting toxicity from known or unknown samples.

Approach

A five-step design will be used to address the study objectives:

1) Create a Governance structure
2) Analyses of historical data and lab techniques provided by ELAP-accredited laboratories to identify sources of variability
3) Dose-response testing to optimize lab technique(s) and recommended lab technique guidance
4) Evaluation of the revised lab technique guidance via split-sample testing by accredited laboratories
5) Final report with final recommended guidance

These steps are sequential with each one informing the details of the next.

The first step will create a two-tiered governance structure to ensure transparency and technical rigor. One tier will be a Stakeholder Committee comprised of representatives from sectors potentially impacted by the study results. The second tier will be an independent Expert Science Panel comprised of scientists with different areas of expertise and no potential conflict with study results.

The second step will be comprised of two subtasks. The first subtask will create an inventory of lab techniques used by ELAP accredited laboratories. The inventory will elucidate the level of comparability and differences in test implementation. The second subtask will collect historical testing data from the ELAP accredited laboratories. The historical data will be analyzed to quantify the level of variability within and among laboratories. Finally, the differences in lab techniques will be compared to the lab test result variability. The goal of this data analysis is to indicate which lab techniques might be accounting for the observed variability. An optional subtask is to collect new data to assess intra- or interlaboratory variability using split samples, to confirm possible sources of test variation that historical data does not provide. This split sample testing will be dependent on the availability of additional funds.

The third step will focus on dose-response testing procedures to quantify the variability of lab techniques identified by the historical analysis in Step 2. There may potentially be many variable-inducing differences in lab techniques from Step 2. So, a prioritization of which techniques require dose-response testing will need to be evaluated by the Stakeholder Committee and Expert Science Panel formed in Step 1. The results of the dose-response testing will culminate in a draft recommendation on the lab technique guidance that produces the least variability in test results.

The fourth step will verify that draft recommendation from Step 3 does reduce variability, both within and among laboratories. To accomplish this verification, split samples will be distributed to ELAP accredited laboratories for testing, using the draft recommended guidance for laboratories to follow.

The fifth step will complete the study including final recommended guidance, results from the split sample testing in Step 4, and a final report.

Detailed Methods

Create a governance structure

This task will create a multi-tiered governance structure to ensure transparency and technical rigor. Figure 1 illustrates the governance structure.

The ultimate decision-making body is the State Water Resources Control Board (SWRCB). The SWRCB staff is charged with making the final recommendations to the SWRCB about the need for implementing any recommended guidance on lab technique for the *C. dubia* reproduction test.

The project facilitator is the Southern California Coastal Water Research Project (SCCWRP). SCCWRP will be responsible for project design (including this workplan), project implementation (including interacting with ELAP accredited laboratories), and project reporting (including the final report).

One tier will be a Stakeholder Committee comprised of representatives from sectors potentially impacted by the study results. A list of the Stakeholder Committee sectors, and their designated representatives is included in Table 1. The goal of the Stakeholder Committee is to ensure there is a formal mechanism for input and feedback to the project design, planning, conclusions, and recommendations. While not a decision-making body, the Stakeholder Committee is a crucial piece of governance. The Stakeholder Committee will review any study design, results, and recommended guidance first, prior to the Expert Science Panel, to make sure the study is rooted in applicable and achievable guidance.

The second tier will be an independent Expert Science Panel comprised of scientists with different areas of expertise and no potential conflict with study results. A list of the Expert Science Panel disciplines and the designated scientists are listed in Table 2. The Expert Science Panel is a decision-making body and is tasked with reviewing the study design and approving the Workplan, reviewing intermediate work products and refining the study design, and reviewing the recommended changes to lab techniques and providing a consensus opinion on the final method guidance.

Figure 1. Project governance structure

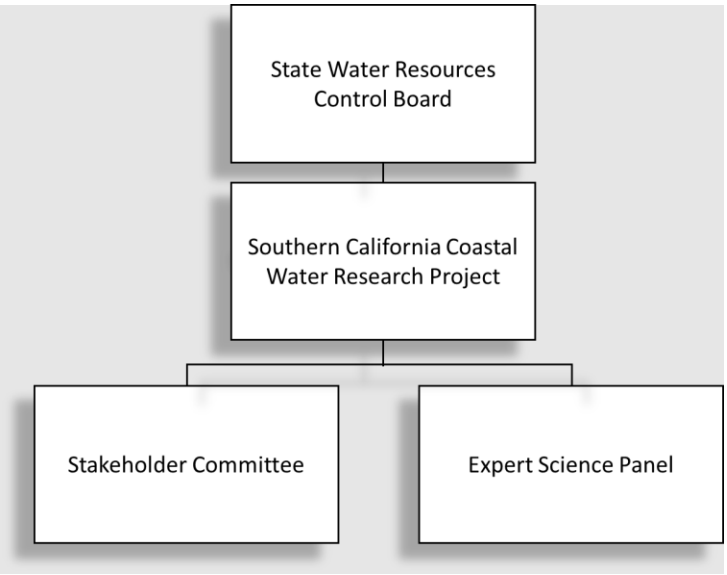


Table 1. List of Stakeholder Committee sectors and designated representatives

SECTOR	**Representative**
Local Government	John Wheeler (SWRCB)
Federal Government	Debra Denton (EPA Region IX)
Regional Water Board; NPDES Permitting	Veronica Cuevas (RWQCB4)
Wastewater Agencies	Mitch Mysliwiec (LWA representing CASA)
Stormwater Agencies	Jian Peng (CASQA)
Agricultural Coalition	Sarah Lopez (Central Coast, CCWQP)
Non-Governmental Organization	Kaitlyn Kalua (CA Coastkeeper)
Private Laboratories	Jeff Miller (Aqua-Science Laboratories)
Public Laboratories	Josh Westfall (Los Angeles County Sanitation Districts)

Table 2. List of Expert Science Panel disciplines and designated scientists.

Expertise	**Representative**
Freshwater Toxicology- Academic	Robert Brent (James Madison University)
Freshwater Toxicology- Government	Teresa Norberg-King (USEPA)
Freshwater Toxicology- Industry	Howard Bailey (Nautilus Environmental)
Biostatistics	A. John Bailer (Miami University)
Data Quality Objectives for WET testing	Leana Van der Vliet (Environment Canada)

Inventory of lab techniques and analyses of historical data

Inventory of Laboratory techniques

To date, no one has created an inventory of laboratory techniques used by ELAP accredited laboratories for the *C. dubia* reproduction test. This task will create that inventory, some of which will be easily accessible and expected based on the promulgated method, and some of which is expected to be more difficult. The inventory will focus on four types of lab techniques including:

- Dilution water
- Food
- Culturing
- Technician training and laboratory's level of experience

Table 3 distills the factors within each of the four types of data to be inventoried, and how they will be collected. For the most part, these categories correspond to parts of the promulgated method that allows for flexibility in their lab techniques.

Multiple approaches will be used to compile the lab technique information including Standard Operating Procedures (SOPs), supporting documents such as bench sheets and quality assurance plans, and a survey questionnaire. The questionnaire will be created after reviewing SOPs, ensuring that all of the relevant information can be collected. If necessary, follow-up one-on-one interviews with lab managers or lab directors may be conducted to verify lab techniques and fill in any missing information.

Compile historical test data

Historical testing data will be compiled from ELAP accredited laboratories. Testing data will focus on controls and each laboratory's reference toxicant results. Test acceptability requirements for controls dictate minimum lab performance and, because there is a lack of toxicant exposure, it is assumed that this represents each lab performing the test to the best of their ability. Similarly, reference toxicant test requirements dictate specified lab performance, especially in precision of test organism response. Test sample data response will not be utilized because test performance expectations are not known.

The goal is to compile the daily number of neonates per replicate for each test and supporting data. Table 3 lists the factors to be calculated from compiling this historical test data including:

- Average number of neonates/female
- Average number of broods/female
- Age of females at test start
- Control variability as standard deviation (SD), coefficient of variation (CV)
- Reference toxicant 50% lethal concentration (LC50) variability as SD and CV
- Reference toxicant 25% inhibitory concentration (IC25 for reproduction) variability as SD and CV
- Test water quality data
- Number of replicates tested

In addition, laboratories may be queried for testing details including:

- Procedure for determining mortality
- Procedure to exclude 4th broods
- Frequency of test failures (if data not provided)

Table 3. Categories of information types to be collected and their likely method of collection.

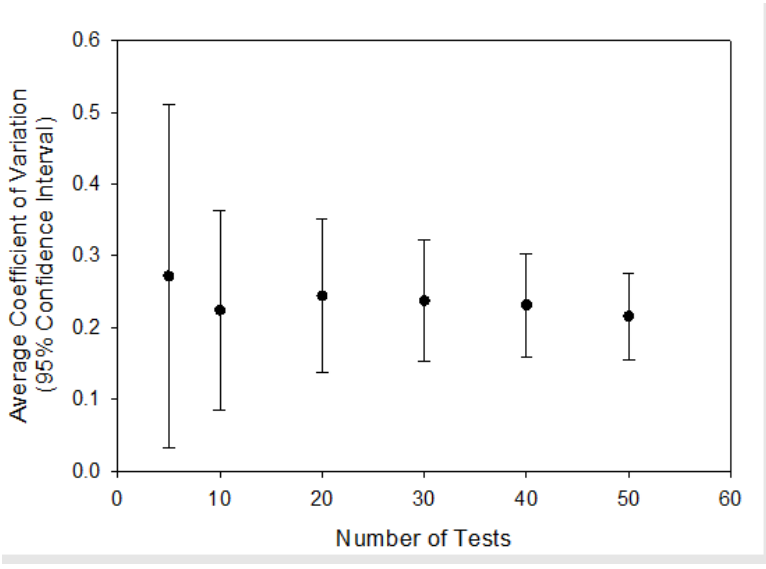
	SOP and supporting documents (e.g. QAP, bench sheets)	Survey and/or phone call	CETIS report and raw data
Dilution water			
Recipe (incl. supplements), vendor	x	x	
Source water	x		
Shelf-time	x	x	
Food			
YCT recipe , vendor	x	x	
Shelf-time	x		
Algal species, source, culture media	x	x	
C. dubia culture			
Frequency of restart/turnover		x	
Frequency of culture failure		x	
Photoperiod	x		
Culture water quality data (e.g. hardness)	x		
Testing procedure/ historical data			
Control variability			x
Age of females at test initiation	x		x
# of neonates/female in controls			x
# of broods/female in controls			x
Daily number of neonates			x
Reference toxicant variability			x
Frequency of test failures	x		x
Test water quality data		x	x
Number of replicates			
Treatment of data outlier		x	
Procedure to exclude 4th broods	x	x	
Time to reproduction	x		
Technician Experience			
Training protocols		x	
Technical experience	x		

Sample size for the historical data compilation is a critical study element. Currently, there are 17 ELAP accredited laboratories conducting the *C. dubia* test in California. Thus, sample size to compare differences among labs is already truncated. The study goal is to collect data from every ELAP accredited laboratory.

Sample size for within laboratory variability assessment is also critical. The study goal is to compile at least 30 tests or the last 3 years of test data from every ELAP accredited laboratory, whichever comes first. Knowing that laboratories do not have the same testing frequency, this target sample size is based on weighing two competing factors that could influence variability assessments. The first factor is the desire to have as much data as possible to have confidence in the variability quantification. The second factor is the desire to keep data as current as possible to reduce the effects of potentially variable-inducing parameters such as evolving lab techniques, turnover in personnel, and other challenges.

Figure 2 presents the results of a simulation to assess confidence in control variability at various sample sizes. This simulation focused on the coefficient of variation (CV) in control brood size per female. Lower CVs have less variability than larger CVs, and Fox et al (2019) targets CVs<0.25 as preferred for reducing errors using the TST. For this simulation, the number of neonates per female ranged from 15 – 30, which meets the promulgated method minimum and approximates the example used in USEPA (2012). The number of neonates per female was randomly assigned for 10 replicates per test and the control CV calculated. This was repeated for sample sizes ranging from 5 to 50 tests. The average and 95% confidence interval of brood size CV per test was calculated for varying sample sizes in Figure 2. This simulation illustrates two important points; a) the average CV tends to stabilize after about 20 tests, and b) the 95% confidence interval about the average CV continues to get smaller with more tests (as expected), but large gains in confidence subside after sample sizes >30. These two results support the study goal of 30 tests or 3 years, whichever comes first.

Figure 2. Simulation to assess average and 95% confidence intervals in control coefficient of variation (CV) at various sample sizes.



Data management, quality assurance and analysis

Data workflow will follow the schema in Figure 3. Input data will come in two categories mirroring the subtasks for this step: laboratory techniques and historical data.

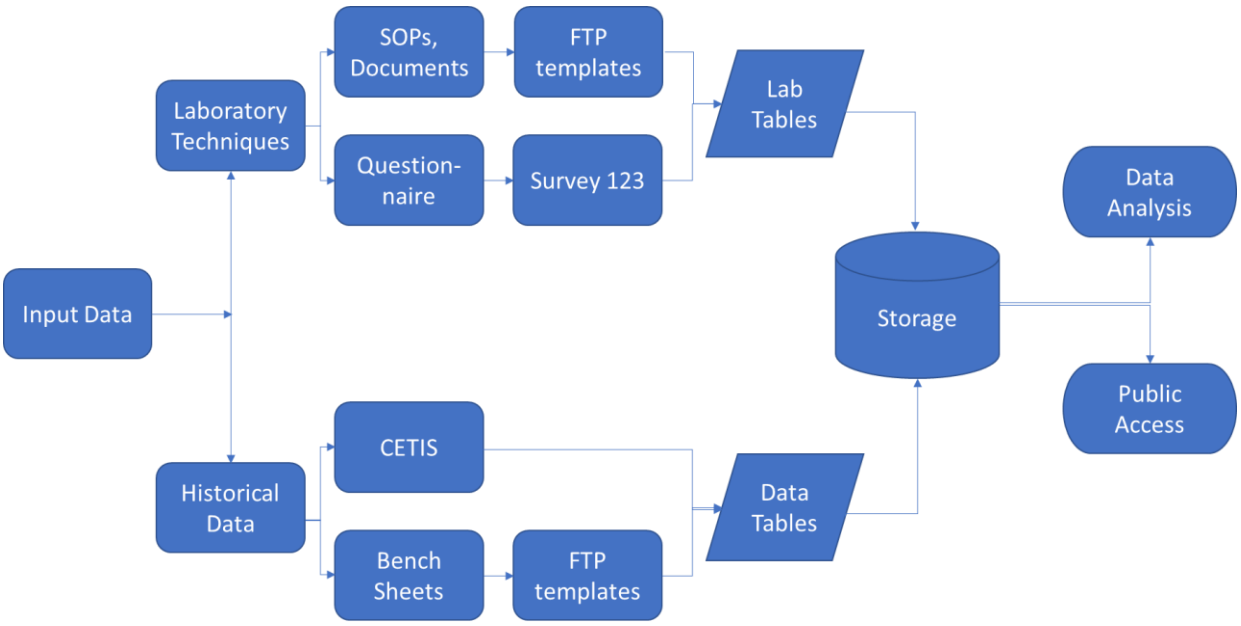
Laboratory technique data will be gathered in two different formats. The first format will be document extraction. This includes documents such as SOPs, Quality Assurance Plans, and other supporting documentation. These data will be categorical, and input using file transfer protocol (ftp) data entry templates by SCCWRP staff. At least 20% of the hand data entered will be checked manually by supervisors and, if errors are found, the remaining 80% of data from that data entry personnel will be checked and accompanied by additional personnel training. The ftp process will also include automated data checkers prior to storage for a 100% data audit of completeness, redundancy, syntax, and look up list errors. The second format will be the laboratory survey, where directed questions will be asked of lab managers or lab directors. The electronic survey will be created in Survey 123, with pre-defined look up lists to eliminate any data entry errors. After automated data checks for completeness and redundancy, these data will be loaded into storage.

Historical data will be more data management intensive than laboratory technique data. The preferred route for compiling historical data is to utilize CETIS (Comprehensive Environmental Toxicology Information System), a software package utilized by regulated agencies for submitting compliance toxicology data. CETIS provides all of the data necessary for this project including daily counts of surviving females and neonate production per replicate, as well as water quality monitoring. SCCWRP will create an ftp for receiving exported CETIS files and transforming into the necessary formats for this project. Automated data checkers prior to storage will create a 100% data audit of completeness, redundancy, syntax, and look up list errors. Where errors occur in exported CETIS data, SCCWRP will query the laboratory of origin for the missing data.

For historical data that is not in CETIS, SCCWRP will need to hand enter data directly from bench sheets. Similar to the workflow for laboratory techniques, these continuous data will utilize file transfer protocol (ftp) data entry templates input by SCCWRP staff. At least 20% of the hand data entered will be checked manually by supervisors and, if errors are found, the remaining 80% of data from that data entry personnel will be checked and accompanied by additional personnel training. The ftp process will also include automated data checkers prior to storage for a 100% data audit of completeness, redundancy, range checks, syntax, and look up list errors.

Once completed, the central data storage will link the laboratory technique and historical data through unique key fields to bind the tests specific to each laboratory. All laboratories will remain anonymous for this study to enhance laboratory participation. The central data storage will be located behind firewalls on SCCWRP servers and backed up twice daily in at least two locations to ensure data security. The laboratory technique and historical data will become publicly accessible data at the conclusion of the study. However, all laboratory identifiers will be kept anonymous to ensure that laboratories are free to participate without fear of being singled out.

Figure 3. Data workflow for this study.



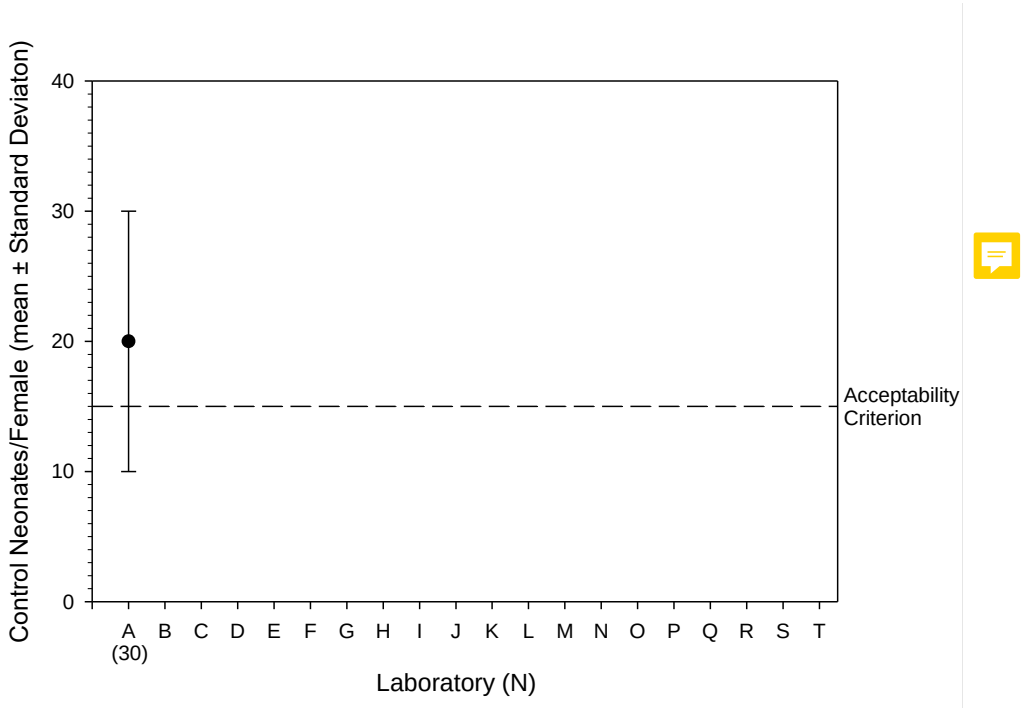
Data analysis will focus on three aspects for this task, consistent with the study questions:

1. Lab technique inventory
2. Comparison of lab performance among laboratories based on historical data
3. Relationships between lab performance and historical data

The lab technique inventory will be a table that quantifies the number of labs utilizing each technique.

Comparison of lab performance among laboratories will utilize pairwise or multiple pairwise (e.g., ANOVA) analysis between laboratories for key test performance metrics. Key test performance metrics will focus on the number of neonates per female (average, standard deviation, CV) in controls and reference toxicant effect concentrations (LC50 and IC25, average, standard deviation, CV). An example of one keystone graphic to compare lab performance among laboratories is Figure 4. To be clear, while statistical analysis will be applied to these data, statistically significant differences are not the goal of this analysis. Especially with so few labs and expected unequal sample sizes among labs, it is anticipated that patterns will be just as important as statistical testing. Where patterns emerge, the lab techniques for the best and worst performing laboratories can be more carefully examined.

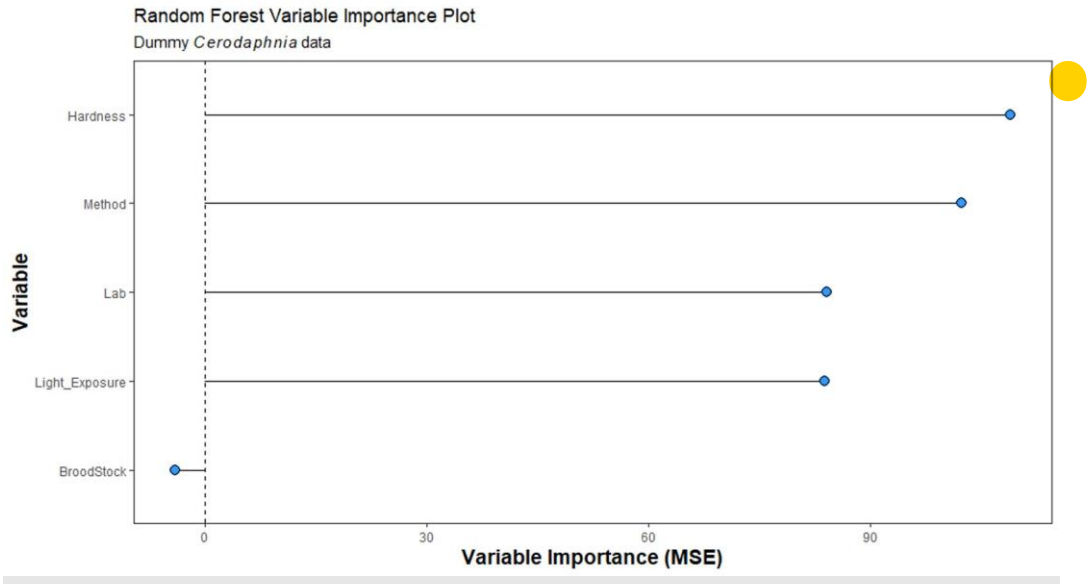
Figure 4. Example keystone graphic for comparing control performance among laboratories. Data for Lab A is not real and plotted only as an example. Data is not plotted for Labs B-T.



Relationships between lab techniques and historical data is the ultimate goal of this task. Multivariate statistics will be used to compare categories and subcategories of lab techniques and the historical data test metrics to determine which lab technique(s) is the source of the most variability (e.g., dilution water recipe versus food source, versus laboratory experience, etc.). The multivariate analysis will include Random Forest and/or Linear Mixed Effects Models, which will attempt to identify variables of greatest importance. An example keystone graphic may look like Figure 5, where mean square error is used to rank variable importance. The data in figure 5 are not real and should be used only for visualization of example end product.

From this series of analyses, a set of proposed revisions or standardizations to the laboratory techniques will be developed. If it is determined that it would be useful to gather more data from the laboratories to verify what was established during the data analysis, then round-robin testing may be proposed if there is sufficient funding. This exercise would seek the participation of all ELAP accredited laboratories to analyze select split samples (e.g., dilution water with different hardness).

Figure 5. Example of keystone graphic illustrating results of random forest. Data is not real, and is for illustration purposes only.



Dose-response testing to optimize lab techniques

The third of five steps in this study will focus on dose-response testing procedures to quantify the variability of lab techniques identified by the historical analysis in Step 2.

There may potentially be many variable inducing differences in lab techniques from Step 2, and not all of them may be followed up in this Step 3. So, a prioritization of which lab techniques require dose-response testing will need to be evaluated by the Stakeholder Committee and Expert Science Panel formed in Step 1. The criteria for prioritization should include:

- Lab techniques that appear to contribute the greatest within or among test variability
- Lab techniques that have multiple options across the most laboratories
- Techniques that have optional approaches in the promulgated method guidance
- Techniques that are not defined in the promulgated method guidance
- Others as agreed upon by the Stakeholder Committee

The dose-response testing will be conducted by one (or a limited number) of laboratories to remove confounding interlaboratory variability. This should help to isolate the variability associated with only the lab technique. Therefore, the laboratory(ies) selected should be amongst the most competent and experienced to ensure capability, minimize intra-laboratory variability, and need not be ELAP accredited.

The results of the dose-response testing will culminate in a technical memorandum that describes the variations of the lab technique selected to be tested, the results of the dose-response testing, and a draft recommended guidance on the optimized lab technique to produce the least variability in test results.

Evaluation of the revised lab technique

The fourth of five steps in this study will be for all of the ELAP accredited laboratories to participate in a round-robin split-sample exercise using the draft laboratory technique guidance developed from Step 3.

Since the project builds from step-to-step, the exact sample types, number of laboratories, and the laboratory technique guidance are currently unknown. The sample types will be chosen so that the effect of the draft laboratory technique guidance can best be tested and quantified. It is presumed that the challenge samples may include a variety of blank samples, as well as some spiked samples. However, the exact number of samples will be agreed upon after Step 3. The number of laboratories will be dependent upon the amount of additional resources available.

Whatever criteria are used to select samples for testing, they first will be vetted by the Stakeholder Committee and then approved by the Expert Science Panel.

The results of the split-sample testing will culminate in a technical memorandum that describes the split-samples created and distributed to laboratories, the results of the split-sample testing, and an assessment of the final draft recommended guidance on the optimized lab technique to produce the least variability in test results.

Final report with final recommended guidance

The final report will summarize the study objectives, methods, results, and a discussion of the findings and limitations of the study. The final report will include the interim deliverables contained within the Technical Memos from Steps 2-4. The final report will also be published documentation to accompany the project database.

Most importantly, the final report will contain the vetted recommended guidance for laboratory activities to optimize variability implementing the *C. dubia* reproduction test.

The Stakeholder Committee will have multiple opportunities to review and provide input on the final report. The Expert Science Panel will also review the final report and provide a consensus opinion on the recommended laboratory technique guidance for implementing the *C. dubia* reproduction test.

SWRCB staff will be responsible for deciding the final disposition of the recommended laboratory technique guidance, and the final recommendation to the State Water Board.

Schedule

Task	Product	Deadline
Study Workplan		
Draft	Draft Workplan to identify potentially variable-inducing lab techniques	3/1/21
Final	Final Workplan approved by Expert Science Panel	5/1/21
Historical Data Analysis		
Lab Data Analysis	Technical Memo identifying potentially variable-inducing lab techniques	7/1/21
Split Sample Analysis (if conducted)	Technical Memo quantifying within and among lab variability	1/1/22
Optimization Testing	Technical memo with draft recommended guidance to reduce within and among lab variability	3/30/22
Interlaboratory Testing	Technical Memo quantifying within and among lab variability	7/31/22
Final Report		
Draft	Draft Report with final recommended guidance	11/1/22
Final	Final Report approved by Expert Science Panel	12/31/22

References

Diamond, J., P. Stribling, M. Bowersox, and H. Latimer. 2008. Evaluation of effluent toxicity as an indicator of aquatic life condition in effluent dominated streams: A pilot study. *Integrated Environmental Assessment and Management* 4:456-470.

Fox, J.F., D.L. Denton, J. Diamond, and R. Stuber. 2019. Comparison of false-positive rates of 2 hypothesis-test approaches in relation to laboratory toxicity test performance. *Environmental Toxiciology and Chemistry* 38:511-523.

Moore, T.F., S.P. Canton, and M. Grimes. 2000. Investigating the incidence of type I errors for chronic whole effluent toxcitiy testing using *Ceriodaphnia dubia*. *Environmental Toxiciology and Chemistry* 19:118-122.

Schiff, K.C. and D. Greenstein. 2016. Stormwater monitoring coalition toxcity testing laboratory guidance document. Southern California Coastal Water Research Project. Costa Mesa, CA.

USEPA. 2001. Final report: Interlaboratory variability study of EPA short-term chronic and acute whole effluent toxicity test methods, Vol. 1. U.S. Environmental Protection Agency. Washington, DC.

USEPA. 2002. Short-term methods for estimating the chronic toxicity of effluents and receiving water to freshwater organisms. EPA-821-R02-013. U.S. Envirnmental Protection Agency. Washington DC.

APPENDIX A – List of ELAP Accredited Laboratories

Lab Name	**Lab Type**
ELAP accredited laboratories in California	
49er Water Laboratory	Private
Aqua-Science	Private
Aquatic Bioassay & Consulting Laboratories, Inc.	Private
Aquatic Testing Laboratories	Private
Aquatic Toxicology Laboratory, Aquatic Health Program	Academic
Enthalpy Analytical, LLC (Nautilus)	Private
Environmental Monitoring Div. (EMD) Lab. at Hyperion Treatment Plant	Public
Granite Canyon -- UC Davis Marine Pollution Studies Laboratory	Academic
Inland Empire Utilities Agency Laboratory	Public
MBC Aquatic Sciences	Private
McCampbell Analytical, Inc.	Private
Pacific EcoRisk	Private
San Jose Creek Water Quality Laboratory	Public
Wood Environment & Infrastructure Solutions, Inc.	Private
ELAP accredited laboratories outside of California	
Eurofins TestAmerica - Corvallis (ASL)	Private
GEI Consultants, Inc.	Private
Ramboll	Private
Tetra Tech's Ecological Testing Facility	Private

Hi Alvina

I have a few brief comments on the SCCWRP Draft Conceptual Workplan: Quality Assurance Evaluation of the *C. dubia* Reproduction Test. Additional, more detailed comments will be provided when the revised Draft Workplan is available from the Expert Science Panel.

1. We agreed that a detailed QAPP should be prepared for the study. This is going to be a major task that will require significant effort so preparation should begin as soon as possible after the proposed Workplan from the Expert Science Panel is finalized.

2. Data generated in this study should be of high quality appropriate for a scientific publication. Who will be responsible for preparing the manuscript from the study data?

3. It is unclear the value of the dose-response testing, particularly with one or a few labs. Perhaps the Expert Science Panel can provide some guidance/justification on this issue.

Thank you for the opportunity to comment.

Jeff Miller, Ph.D. DABT
President
AQUA-Science, LLC
630 Cantrill Dr
Davis, CA 95618
(530)753-5456
aqua-science.com

Thank you for considering the wastewater sector comments on the Draft Conceptual Workplan (SCCWRP), dated March 19, 2021, for the State Water Resources Control Board (State Water Board or SWB) *Ceriodaphnia dubia* study.

1) The study schedule developed by the SWB[1] described a "draft Study Work Plan" and a "final Study Work Plan". The draft would be reviewed by an Expert Science Panel (Science Panel) and the Stakeholder Advisory Committee (SAC) so that a final Study Work Plan would be completed by March 2021. The current document being reviewed by the SAC is called a "Draft Conceptual Workplan". This Draft Conceptual Workplan is lacking in many details that are necessary to clearly understand the scope of the study and how it will be implemented. While we understand the intent is to create the Draft Study Work Plan after receiving Science Panel comments, providing more detail at this stage may provide more for the Science Panel to consider and comment on. While the Draft Study Work Plan will receive review by the Science Panel, the timing of the project is slipping and efforts to facilitate the review should be considered. A revised/detailed workplan was mentioned at the SAC meeting on 3/24/21, but it is not clear if or how the Draft Conceptual Workplan and a revised workplan are replacing the Study Work Plan identified by the SWB. Additionally, it is not clear how these documents fit into the study schedule. Please consider creating a process outline for the website with associated dates updated as necessary. It would be helpful to clarify the following:
    a. Explain if the Conceptual Workplan and a revised workplan are intended to replace the draft and final workplans desribed in the SWB study schedule.
    b. Describe what information would be included or expanded in the revised/detailed workplan and how is it contingent on information not yet known.
    c. Update the schedule (page 14) to indicate when the revised workplan will be developed and shared for review.

2) There are two objectives written in the Draft Conceptual Workplan, paraphrasing them they would be: a) investigate and improve intra-lab variability, and b) improve inter-lab variability. However, in the subsequent study questions, there does not appear to be a question directed at improving inter-lab variability. Consider adding, "…between laboratories" to the end of question 3.

3) Please identify the study questions being addressed in tasks 2 through 4.

4) The Draft Conceptual Workplan does not provide sufficient details necessary to clearly implement the study – even the immediate next step (i.e., Step 2). Several members of the SAC meeting on 3/24/21 also commented that more information was needed. While all information needed to describe Step 3 and Step 4 may not be available until Steps 1 and 2 are complete, it is appropriate to include as much information and detail as possible at this time. Specific details recommended to describe in this workplan include:

[1] https://www.waterboards.ca.gov/water_issues/programs/state_implementation_policy/tx_ass_cntrl.html

a. Describe Data Quality Objectives (DQOs) according to USEPA (2006) guidance, as recommended by a member of the SAC. The DQOs would identify specific study questions (e.g., testable hypotheses) to be answered. Specific study questions can be developed at this time and should be linked to each of the data needs so the Science Panel will know what questions the study is specifically trying to answer and if the proposed approach workplan will be able to answer them.
b. Table 3 provides only categories of information and likely methods of collection, but this is not enough detail to understand if the information in mind will meet the goals, objectives, and questions of the study. It would be very helpful to list specific information and data that will be requested from labs for analysis. For example,
    i. Are basic water quality parameters (pH, Hardness, ion balance, conductivity, etc) included in dilution water recipe? Consider adding water quality parameters as a specific call out.
    ii. What specific data will be requested under the category of control variability? Is this only CVs or does it include lab control survival and reproduction, individual neonates counts, etc.?
    iii. What specific reference toxicity data will be evaluated? (e.g., individual test CVs, IC25s, IC10s, standard deviations)
    iv. Under testing procedure consider asking survey questions:
        1. What modifications have been made to improve CVs
        2. What modifications were tried but abandoned as ineffective?
    v. How will technical experience be determined/quantified/analyzed and what metrics will be used to describe it?
    vi. Data describing the specific age of neonates at the start of testing (e.g., 0-8 hrs, 8-16 hrs, or 16-24 hrs old) should be requested.

5) The *excerpt from Agreement Number 19-078-270 (The portions relevant to the C. dubia study)* states:

   "The Study Work Plan shall detail, at a minimum, the study approach to evaluate historical performance data and other metrics to assess within-laboratory variability and a design for the within-laboratory and between-laboratory comparability testing."

   However, the Draft Conceptual Workplan identifies between-laboratory comparability testing as an "optional subtask."

   "An optional subtask is to collect new data to assess intra- or interlaboratory variability using split samples, to confirm possible sources of test variation that historical data does not provide. This split sample testing will be dependent on the availability of additional funds."

   The study goals should list the intent to identify and reduce sources of interlaboratory variability in *C. dubia* test results to be consistent with the SWB *C. dubia* study description.

6) The timing for initiating intra- or inter-laboratory variability testing should be clearly identified. As stated in the Draft Conceptual Workplan, the "…optional subtask is to collect new data to assess intra- or interlaboratory variability using split samples…" is described in Step 2 of the

study. The rationale for conducting this testing as part of Step 2 was "…to confirm possible sources of test variation that historical data does not provide." However, at the 3/24/21 meeting a SAC member suggested that this testing should occur in Step 3 or Step 4. A subsequent action item from the meeting was to ask the Science Panel to advise on the timing for such testing. Testing to evaluate inter-laboratory variability may be appropriate for Steps 2, 3, and 4. If there are options for the timing of inter-laboratory variability testing (e.g., if not necessary to conduct as part of Step 2 as described) then it would be helpful to describe alternative schedules and how such data would or would not support the study goals.

7) The rationale for selecting one (or a limited number) of labs for controlled variable "dose-response" testing (Step 3) should be further explained and technically justified. The Draft Conceptual Workplan states that only the "most competent and experienced" lab(s) will be selected for this testing, although the qualifications are undefined (e.g., a lab with low CV?). A lab with low control CVs has less room for improvement and less need to lower CVs. Therefore, this approach seems to be of limited utility because conducting controlled-variable testing with only the most competent lab(s) would bias the outcome of the study. Only factors that result in relatively small improvements to CVs/performance would be identified (because there is not much room for improvement) and factors with a large influence on performance at poorer performing labs, might not be identified. Also, an underlying assumption seems to be that revising methods at a high-performing lab will cause the same result at a different laboratory. This assumption should be explicitly stated and vetted by the Science Panel. Please clarify the rationale for lab selection in Step 3 and describe how interlaboratory variability will either be incorporated or why it will be excluded from testing in Step 3.

8) The Draft Conceptual Workplan proposes to compile historical test data over the past 3 years or 30 tests (page 8). This approach may exclude helpful information labs have developed to assess and reduce test performance (i.e., lower control CVs) prior to this period. A rationale is provided for the proposed period, but compiling information from labs outside of this period could lead to a more successful study if it includes data identifying sources of variability in culture health or testing or lessons learned and what the lab did to improve performance. We recommend describing how potentially useful information (qualitative and/or quantitative) from beyond this relatively short period will be solicited and utilized in the *C. dubia* study.

March 29, 2021

Ken Schiff
Southern California Coastal Water Research Project (SCCWRP)
3535 Harbor Blvd., Suite 10
Costa Mesa, CA 92626

Re: *Ceriodaphnia dubia* Toxicity Testing Quality Assurance: Work Plan

Dear Mr. Schiff,

The California Coastkeeper Alliance (CCKA) represents local California Waterkeeper organizations working to protect and enhance clean waters throughout the state for the benefit of Californians and California ecosystems. CCKA participated in the State Water Resources Control Board's (State Water Board) development and final adoption of the Toxicity Provisions, and remains committed to its timely implementation of numeric limits for acute and chronic toxicity. On behalf of local Waterkeepers, we appreciate the opportunity to comment on the draft work plan for the *Ceriodaphnia dubia* Toxicity Testing Quality Assurance study and the related recommendations provided by the California Association of Sanitation Agencies (CASA).

The *Ceriodaphnia dubia* Toxicity Testing Quality Assurance study resulted from a perceived lack of reliability and lack of confidence in the *ceriodaphnia dubia* chronic toxicity test that was raised during the public process for the final Toxicity Provisions. The State Water Board observed that while most laboratories can achieve the desired performance level, however, there is variation in laboratory performance in California. For this reason, the State Water Board adopted the final Toxicity Provisions with the direction that a study be conducted to evaluate whether best practices and guidelines may be recommended to California laboratories to improve laboratory performance statewide, and support the use of *ceriodaphnia dubia* for regulatory compliance in water quality permits no later than January 1, 2024.

It is with this context we provide the following comments regarding CASA's recommendations for the study work plan. Specifically, we urge the Science Expert Panel be instructed to independently review and either include or omit CASA's recommendations, and ensure that recommendations regarding the study design are not be contingent on funding provided by Stakeholder Advisory Committee members – members who already have the opportunity to oversee and raise questions regarding the design and questions addressed by the study.

A. CASA's recommendations should be independently evaluated by the Science Expert Panel for consistency and relevance to the current study scope.

As described by the State Water Board staff during the first Stakeholder Advisory Committee meeting this past December, the study "is a quality assurance study to determine whether laboratory best practices might be recommended to improve laboratory performance,"[1] *not* a method validation study to determine whether *ceriodaphnia dubia* should be used in California regulatory programs nor a study to estimate false positive or false negative rates using the test of Significant Toxicity (TST). CASA's recommendations must be reviewed through the lens of whether the recommendations are within or outside the study scope, as directed by the State Water Board.

[1] Stakeholder Advisory Committee Meeting, December 8, 2020.

B. The evaluation and potential inclusion of CASA's recommendations must avoid direct conflict of interest and funding bias.

The study must remain objective and maintain high standards of scientific integrity, including the elimination of funding or sponsorship bias, and conflicts of interest be minimized. Scientific studies may be subject to implicit bias caused indirectly by the funding source, however, CASA's recommendations include an alarming and explicit funding bias and conflict of interest:

> "The scope of CASA and the POTW community funding will ultimately be contingent on how well the study design meets stakeholder concerns and interests identified above."[2]

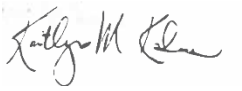
This condition is unacceptable and should be struck from recommendations provided to the Science Expert Panel with clear direction that the Panel shall independently review and either incorporate or omit the recommendations based on their technical expertise and best scientific judgment.

It is critical that the final study design be objectively and independently evaluated by Science Expert Panel – and not unduly influenced by the funding source of the study. The Science Expert Panel must have the independent discretion to provide recommendations and expert findings regarding the design of the study, whether laboratory testing is necessary for the scope of purpose of the study, and what factors shall be evaluated and measured. As was mentioned during the third Stakeholder Advisory Committee meeting held on March 24, 2021, evaluating the factors outlined in CASA's recommendations is also premature, given the evaluation of historical data and resulting recommendations has not yet been completed. Further, CASA's recommendations inappropriately state the recommendations are provided as "expectations"[3] from CASA and other stakeholders, and should instead be framed as recommendations for the Science Expert Panel to ***consider*** – not be mandated.

We request that CASA's recommendations only be provided to the Science Expert Panel with the following instructions:

- That the recommendations be independently reviewed and only incorporated, in whole or in part, based on the technical expertise and opinion of the Science Expert Panel members that (1) the recommendations are relevant to the study scope and purpose, and (2) provide scientific benefit to the study scope given contractual requirements and timing of the study.
- The study design shall not be contingent upon the approval by third-party funders.

Sincerely,



Kaitlyn Kalua
Policy Manager
California Coastkeeper Alliance

[2] California Association of Sanitation Agencies (CASA), Recommendations for Laboratory Testing as Part of the State Water Board *Ceriodaphnia* Study, p. 1.
[3] *Id*. at p. 4.